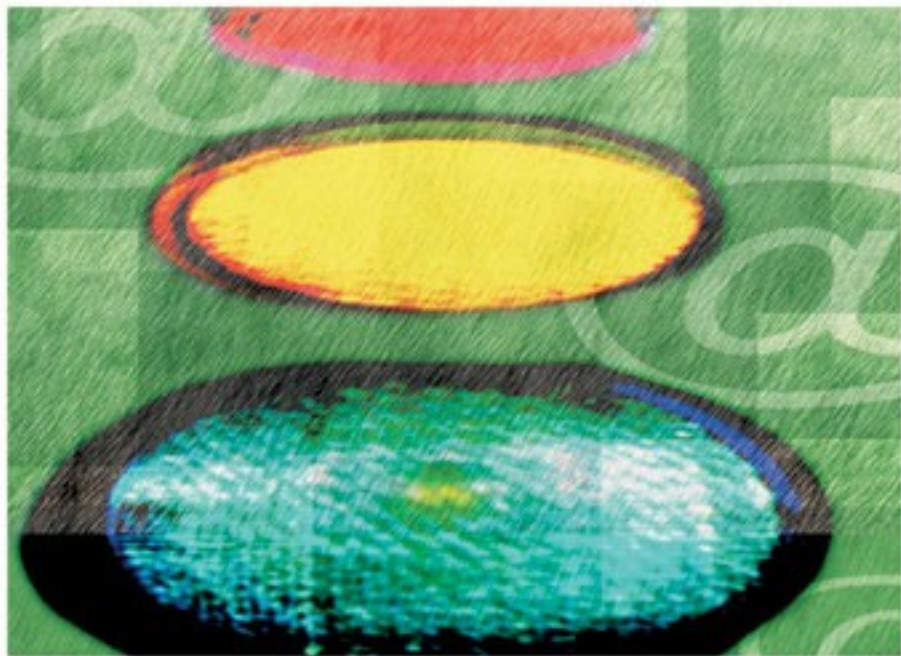


HANDBOOK OF RESEARCH ON

Mobile Business

Technical, Methodological
and Social Perspectives



Bhuvan Unhelkar

Handbook of Research in Mobile Business: Technical, Methodological, and Social Perspectives

Volume I
Chapters 1–30

Bhuvan Unhelkar
University of Western Sydney, Australia



IDEA GROUP REFERENCE
Hershey • London • Melbourne • Singapore

Acquisitions Editor: Michelle Potter
Development Editor: Kristin Roth
Senior Managing Editor: Amanda Appicello
Managing Editor: Jennifer Neidig
Copy Editor: Maria Boyer
Typesetter: Diane Huskinson
Support Staff: Renée Davies
Cover Design: Lisa Tosheff
Printed at: Yurchak Printing Inc.

Published in the United States of America by
Idea Group Reference (an imprint of Idea Group Inc.)
701 E. Chocolate Avenue, Suite 200
Hershey PA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@idea-group.com
Web site: <http://www.idea-group-ref.com>

and in the United Kingdom by
Idea Group Reference (an imprint of Idea Group Inc.)
3 Henrietta Street
Covent Garden
London WC2E 8LU
Tel: 44 20 7240 0856
Fax: 44 20 7379 0609
Web site: <http://www.eurospanonline.com>

Copyright © 2006 by Idea Group Inc. All rights reserved. No part of this publication may be reproduced, stored or distributed in any form or by any means, electronic or mechanical, including photocopying, without written permission from the publisher.

Product or company names used in this set are for identification purposes only. Inclusion of the names of the products or companies does not indicate a claim of ownership by IGI of the trademark or registered trademark.

Library of Congress Cataloging-in-Publication Data

Handbook of research in mobile business : technical, methodological and social perspectives / Bhuvan Unhelkar, editor.
p. cm.

Summary: "This reference book brings together various perspectives on the usage and application of mobile technologies and networks in global business"--Provided by publisher.

Includes bibliographical references and index.

ISBN 1-59140-817-2 (hardcover) -- ISBN 1-59140-818-0 (ebook)

1. Mobile commerce. 2. Mobile communication systems--Economic aspects. I. Unhelkar, Bhuvan.

HF5548.34.H36 2006

658'.05--dc22

2005032111

British Cataloguing in Publication Data

A Cataloguing in Publication record for this book is available from the British Library.

All work contributed to this book is new, previously-unpublished material. The views expressed in this book are those of the authors, but not necessarily of the publisher.

Editorial Advisory Board

Shivprakash Agrawal, *Gujarat University, India*

Marco Garito, *Cisco Systems, Belfast/Northern Ireland*

Elaine Lawrence, *University of Technology, Sydney, Australia*

San Murugesan, *Southern Cross University, Australia*

Mark Neely, *Austereo Group Ltd., Australia*

Ketan Vanjara, *Microsoft, India*

Upkar Varshney, *Georgia State University, USA*

Saba Zamir, *CRC Press, USA*

List of Contributors

Abood, Christopher / <i>Australian Computer Society, Australia</i>	666
Alag, Harpreet / <i>Agilisys Limited, UK</i>	583
Ammi, Chantal / <i>GET/Institut National des Télécommunications, France</i>	420
Arunatileka, Dinesh / <i>University of Western Sydney, Australia</i>	778
Ayadi, Achraf / <i>GET/Institut National des Télécommunications, France</i>	420
Banakar, R. M. / <i>B.V.B. College of Engineering and Technology, Karnataka, India</i>	184
Barbier, Franck / <i>PauWare Research Group, France</i>	200
Barcelo, Francisco / <i>Technical University of Catalonia, Spain</i>	35
Barjis, Joseph / <i>Georgia Southern University, USA</i>	719
Bhattar, Raghunadh K. / <i>Indian Space Research Organization of Indian Institute Science, India</i>	142
Boulmalf, Mohamed / <i>United Arab Emirates University, UAE</i>	322
Brantner, Matthias / <i>University of Mannheim, Germany</i>	54
Brodt, Torsten / <i>University of St. Gallen, Switzerland</i>	754
Chand, Narottam / <i>Indian Institute of Technology Roorkee, India</i>	132
Chatwin, Chris / <i>University of Sussex, UK</i>	379
Chen, Xiao / <i>Nanjing University of Chinese Medicine, China</i>	487
Ciganek, Andrew P. / <i>University of Wisconsin-Milwaukee, USA</i>	675
Claret-Tournier, Fred / <i>University of Sussex, UK</i>	379
Curry, Joanne Marie / <i>University of Western Sydney, Australia</i>	848
Das, Ritanjan / <i>University of Portsmouth, UK</i>	694
Das(Bit), Sipra / <i>Bengal Engineering and Science University, India</i>	297
Dasgupta, K. S. / <i>Indian Space Research Organization, India</i>	142
El-Masri, Samir / <i>University of Western Sydney, Australia</i>	106, 544
Falcone, Francesco / <i>Digital Business, Italy</i>	444
Feuchtmüller, Hartmut / <i>T-Systems International, Germany</i>	870
Fischer, Thomas / <i>T-Systems International, Germany</i>	870
Gan, Jason / <i>University of Technology, Australia</i>	504
Garito, Marco / <i>Digital Business, Italy</i>	444
Ghanbary, Abbass / <i>University of Western Sydney, Australia</i>	602
Godbole, Nina / <i>CQA, CISA, PMP, CSTE, ITIL (Foundation) Certified Professional Member—Computer Society of India, India</i>	463
Goh, John / <i>Monash University, Australia</i>	216
Gomathy, C. / <i>Deemed University, India</i>	308
Guan, Sheng-Uei / <i>Brunel University, UK</i>	366, 509, 527
Hameurlain, A. / <i>IRIT—Paul Sabatier University, France</i>	267
Harmon, Robert / <i>Portland State University, USA</i>	18
Hawking, Paul / <i>Victoria University, Australia</i>	839

Helmer, Sven / <i>University of London, Birkbeck College, UK</i>	54
Hu, Wen-Chen / <i>University of North Dakota, USA</i>	401
Hürster, Walter / <i>T-Systems International, Germany</i>	870
Islam, Mohammad Mahfuzul / <i>Monash University, Australia</i>	332
Johnstone, Bradley / <i>BK Solutions, Australia</i>	765
Jones, Matthew R. / <i>University of Cambridge, UK</i>	69
Joshi, R. C. / <i>Indian Institute of Technology Roorkee, India</i>	132
Kanne, Carl-Christian / <i>University of Mannheim, Germany</i>	54
Karjaluto, Heikki / <i>University of Oulu, Finland</i>	708
Kuppuswami, Anand / <i>University of Western Sydney, Australia</i>	285
Lan, Yi-chen / <i>University of Western Sydney, Australia</i>	630
Lazarus, Sandra Synthia / <i>University of Sydney, Australia</i>	96
Leary, Thomas / <i>University of North Carolina at Greensboro, USA</i>	233
Lee, Chean / <i>Methodscience.com, Australia</i>	643
Lee, Maria Ruey-Yuan / <i>Shih Chien University, Taiwan</i>	630
Lei, Pouwan / <i>University of Bradford, UK</i>	379, 694
Leow, Chye-Huang / <i>Singapore Polytechnic, Singapore</i>	553
Leppäniemi, Matti / <i>University of Oulu, Finland</i>	708
Li, Feng / <i>University of Newcastle upon Tyne, UK</i>	708
Liu, Wei / <i>Nanyang University of Chinese Medicine, China</i>	487
Mammeri, Z. / <i>IRIT—Paul Sabatier University, France</i>	267
Marmaridis, Ioakim / <i>University of Western Sydney, Australia</i>	563
Marsit, N. / <i>IRIT—Paul Sabatier University, France</i>	267
Martin-Escalona, Israel / <i>Technical Univeristy of Catalona, Spain</i>	35
Martins, Henrique M. G. / <i>University of Cambridge, UK</i>	69
McGregor, Carolyn / <i>University of Western Sydney, Australia</i>	83
Mishra, Shailenara / <i>Dehradun Institute of Technology, India</i>	158, 173
Misra, Manoj / <i>Indian Institute of Technology Roorkee, India</i>	132
Mitra, Sulata / <i>Bengal Engineering and Science University, India</i>	297
Moerkotte, Guido / <i>University of Mannheim, Germany</i>	54
Morvan, F. / <i>Paul Sabatier University, France</i>	267
Murshed, Manzur / <i>Monash University, Australia</i>	332
Nand, Sashi / <i>Rushmore University, Grand Cayman, BWI</i>	393
Nugent, Devon / <i>The University of Queensland, Australia</i>	793, 810
Palsule, V. S. / <i>Indian Space Research Organization, India</i>	142
Patel, Amol / <i>ConvergeLabs Corporation, USA</i>	429
Patel, Keyurkumar J. / <i>Box Hill Institute, Australia</i>	728
Paull, Stephen / <i>Victoria University, Australia</i>	839
Priyatamkumar / <i>B.V.B. College of Engineering and Technology, Karnataka, India</i>	184
Quah, Jon Tong-Seng / <i>Nanyang Technological University, Singapore</i>	553
Raisinghani, Mahesh S. / <i>Texas Woman's University, USA</i>	435
Rajeev, S. / <i>PSG College of Technology, India</i>	613
Ramakrishnan, K. R. / <i>Indian Institute of Science, India</i>	142
Ramamurthy, K. / <i>University of Wisconsin-Milwaukee, USA</i>	675
Ranft, Anne-Marie / <i>University of Technology, Australia</i>	655
Rao, N. Raghavendra / <i>SSN School of Management & Computer Applications, India</i>	828
Reyes, Gina / <i>Victoria University, Australia</i>	839
Romeo, Fabien / <i>PauWare Research Group, France</i>	200

Salam, A. F. / <i>University of North Carolina at Greensboro, USA</i>	233, 859
Salo, Jari / <i>University of Oulu, Finland</i>	708
Schilhavy, Richard / <i>University of North Carolina at Greensboro, USA</i>	859
Shankaranand, B. / <i>National Institute of Technology Surathkal Karnataka, India</i>	184
Shanmugavel, S. / <i>Anna University, India</i>	308
Sharma, Pramod / <i>The University of Queensland, Australia</i>	793, 810
Shuaib, Khaled / <i>United Arab Emirates University, UAE</i>	322
Singh, Nipur / <i>Dehradun Institute of Technology, India</i>	158, 173
Singh, Rahul / <i>University of North Carolina at Greensboro, USA</i>	233
Sinisalo, Jaakko / <i>University of Oulu, Finland</i>	708
Sivanandam, S. N. / <i>PSG College of Technology, India</i>	613
Sreenaath, K. V. / <i>PSG College of Technology, India</i>	613
Su, Hsiang-Ju / <i>Shih Chien University, Taiwan</i>	630
Taniar, David / <i>Monash University, Australia</i>	216
Unni, Ramaprasad / <i>Portland State University, USA</i>	18
Vaghjiani, Khimji / <i>BK Solutions, Australia</i>	765
Vanjara, Ketan / <i>Microsoft Corporation, India</i>	113
Vyas, Amrish / <i>University of Maryland, Baltimore County, USA</i>	1
Wang, Jia Jia / <i>University of Bradford, UK</i>	694
Wang, Fu Lee / <i>City University of Hong Kong, Hong Kong</i>	247
Wiggen, Tom / <i>University of North Dakota, USA</i>	401
Wyld, David C. / <i>Southeastern Louisiana University, USA</i>	740
Yang, Christopher C. / <i>Chinese University of Hong Kong, Hong Kong</i>	247
Yang, Hung-Jen / <i>National Kaohsiung Normal University, Taiwan</i>	401
Yoon, Victoria / <i>University of Maryland, Baltimore County, USA</i>	1
Young, Rupert / <i>University of Sussex, UK</i>	379

Table of Contents

Foreword	xxxv

Preface	xxxvii

Section I Location

Chapter I Information Management in Mobile Environments Using a Location-Aware Intelligent Agent System / <i>Amrish Vyas and Victoria Yoon</i>	1

Chapter II Location-Based Services: Opportunities and Challenges / <i>Ramaprasad Unni and Robert Harmon</i>	18

Chapter III Location Services in Cellular Networks / <i>Israel Martin-Escalona and Francisco Barcelo</i>	35

Chapter IV Ontologies for Location-Based Services / <i>Matthias Brantner, Sven Helmer, Carl-Christian Kanne, and Guido Moerkotte</i>	54

Section II Health

Chapter V Relevance of Mobile Computing in the Field of Medicine / <i>Henrique M. G. Martins and Matthew R. Jones</i>	69

Chapter VI Mobility in Healthcare for Remote Intensive Care Unit Clinical Management / <i>Carolyn McGregor</i>	83

Chapter VII

The Evaluation of Wireless Devices Used by Staff at Westmead Hospital, Sydney / <i>Sandra Synthia Lazarus</i>	96

Chapter VIII

Mobile Comprehensive Emergency System / <i>Samir El-Masri</i>	106

Chapter IX

Application of Mobile Technologies in Healthcare Diagnostics and Administration / <i>Ketan Vanjara</i>	113

Section III Technical

Chapter X

Energy-Efficient Cache Invalidation in Wireless Mobile Environment / <i>R. C. Joshi, Manoj Misra, and Narottam Chand</i>	132

Chapter XI

Review of Wireless Technologies and Generations / <i>Raghunadh K. Bhattar, K. R. Ramakrishnan, K. S. Dasgupta, and V. S. Palsule</i>	142

Chapter XII

Orthogonal Complex Quadrature Phase Shift Keying (OCQPSK) Spreading for 3G W-CDMA Systems / <i>Shailendra Mishra and Nipur Singh</i>	158

Chapter XIII

Code and Time Synchronization of the Cell Search Design Influence on W-CDMA System Performance / <i>Shailendra Mishra</i>	173

Chapter XIV

Turbo Equalizer: A Solution to 4G Mobiles / <i>Priyatamkumar, R. M. Banakar, and B. Shankaranand</i>	184

Chapter XV

Administration of Wireless Software Components / <i>Franck Barbier and Fabien Romeo</i>	200

Chapter XVI

Mobile User Data Mining and Its Applications / <i>John Goh and David Taniar</i>	216

Chapter XVII

Dynamic Matching of Supply and Demand in an M-Commerce Services Marketplace: Using Intelligent Agents and Semantic Technology / <i>Thomas Leary, A. F. Salam, and Rahul Singh</i>	233

Chapter XVIII

Information Delivery for Mobile Business: Architecture for Accessing Large Documents through Mobile Devices / <i>Christopher C. Yang and Fu Lee Wang</i>	247

Chapter XIX

Database Queries in Mobile Environments / <i>N. Marsit, A. Hameurlain, Z. Mammeri, and F. Morvan</i>	267

Section IV Network

Chapter XX

A Nueral Network-Based Mobile Architecture for Mobile Agents / <i>Anand Kuppuswami</i>	285

Chapter XXI

Load Balancing as a Key to Enable Different Services in Cellular Network / <i>Sipra Das(Bit), and Sulata Mitra</i>	297

Chapter XXII

A Novel Fuzzy Scheduler for Mobile Ad Hoc Networks / <i>S. Shanmugavel and C. Gomathy</i>	308

Chapter XXIII

Co-Existence of WLAN and WPAN Communication Systems / <i>Khaled Shuaib and Mohamed Boulmalf</i>	322

Chapter XXIV

Mobility Support Resource Management for Mobile Networks / <i>Mohammad Mahfuzul Islam and Manzur Murshed</i>	332

Section V Security

Chapter XXV

Secure Agent Roaming for Mobile Business / <i>Sheng-Uei Guan</i>	366

Chapter XXVI

Tackling Counterfeiting with a Secure Online Track-and-Trace System / <i>Fred Claret-Tournier, Pouwan Lei, Chris Chatwin, and Rupert Young</i>	379

Chapter XXVII

Developing a Theory of Portable Public Key Infrastructure (PORTABLEPKI) for Mobile Business Security / <i>Sashi Nand</i>	393

Chapter XXVIII

Systems, Handheld Devices, and Payment Methods for Mobile Commerce / <i>Wen-Chen Hu, Tom Wiggen, and Hung-Jen Yang</i>	401

Section VI Strategy

Chapter XXIX

Strategic Perspectives in Mobile Banking: Technology, Value Creation, and Developing Factors / <i>Achraf Ayadi and Chantal Ammi</i>	420

Chapter XXX

Mobile Commerce in Emerging Economies / <i>Amol Patel</i>	429

Chapter XXXI

M-Business: A Global Perspective / <i>Mahesh S. Raisinghani</i>	435

Chapter XXXII

Mobile Strategy Roadmap / <i>Francesco Falcone and Marco Garito</i>	444

Chapter XXXIII

Relating Mobile Computing to Mobile Commerce / <i>Nina Godbole</i>	463

Chapter XXXIV

The Future of Mobile Technologies and Applications in China / <i>Xiao Chen and Wei Liu</i>	487

Section VII Application

Chapter XXXV

Developing Smart Client Mobile Applications / <i>Jason Gan</i>	504

Chapter XXXVI

Ontology-Based Information Retrieval Under a Mobile Business Environment / <i>Sheng-Uei Guan</i>	509

Chapter XXXVII

Intelligent Product Brokering Services / <i>Sheng-Uei Guan</i>	527

Chapter XXXVIII

Understanding Mobile Web Services (MWS) and Their Role in Integrating Mobile Devices / <i>Samir El-Masri</i>	544

Chapter XXXIX

Push-Multicasting to Wireless Devices Using Publish/Subscribe Model / <i>Jon Tong-Seng Quah and Chye-Huang Leow</i>	553

Section VIII Method

Chapter XL

A Methodology for M-Transformation of Small and Medium Enterprises (SMEs) and its Application in Practice Using CBEADS® / *Ioakim (Makis) Marmaridis* 563

Chapter XLI

Business Process Mobility / *Harpreet Singh Alag* 583

Chapter XLII

Evaluation of Mobile Technologies in the Context of Their Applications, Limitations, and Transformation / *Abbass Ghanbary* 602

Chapter XLIII

Policy-Based Mobile Computing / *S. Rajeev, S. N. Sivanandam, and K. V. Sreenaath* 613

Section IX Customer

Chapter XLIV

Investigation of Consumer Behavior in Using Mobile Payment Services—A Case Study of Mobile Recreational Services in Taiwan / *Maria Ruey-Yuan Lee, Yi-chen Lan, and Hsiang-Ju Su* 630

Chapter XLV

Mobile CRM: Reaching, Acquiring, and Retaining Mobility Consumers / *Chean Lee* 643

Chapter XLVI

Factors Influencing Segmentation and Demographics of Mobile-Customers / *Anne-Marie Ranft* 655

Section X Social

Chapter XLVII

Mobile Camera Phones—Dealing with Privacy, Harassment, and Spying/Surveillance Concerns / *Christopher Abood* 666

Chapter XLVIII

Social Context for Mobile Computing Device Adoption and Diffusion: A Proposed Research Model and Key Research Ideas / *Andrew P. Ciganeck and K. Ramamurthy* 675

Chapter XLIX

A Socio-Cultural Analysis of the Present and the Future of the M-Commerce Industry /
Ritanjan Das, Jia Jia Wang, and Pouwan Lei 694

Chapter L

The Mobile Network as a New Medium for Marketing Communications: A Case Study /
Heikki Karjaluo, Matti Leppäniemi, Jari Salo, Jaakko Sinisalo, and Feng Li 708

Chapter LI

Overview and Understanding of Mobile Business in the Age of Communication / *Joseph Barjis* 719

Section XI Case Study

Chapter LII

Successful Implementation of Emerging Communication Technologies in a Mobile-Intense
 Organization: A Case Study of Sydney Airport / *Keyurkumar J. Patel* 728

Chapter LIII

The Next Big RFID Application: Correctly Steering Two Billion Bags a Year Through Today's
 Less-Than-Friendly Skies / *David C. Wyld* 740

Chapter LIV

Identified Customer Requirements in Mobile Video Markets—A Pan-European Case / *Torsten Brodt* 754

Chapter LV

Applying Mobility in the Workforce / *Bradley Johnstone and Khimji Vaghjani* 765

Chapter LVI

Applying Mobile Technologies to Banking Business Processes / *Dinesh Arunatileka* 778

Chapter LVII

Mobile GIS—Challenges and Solutions / *Pramod Sharma and Devon Nugent* 793

Chapter LVIII

Mobile Technologies and Tourism / *Pramod Sharma and Devon Nugent* 810

Chapter LIX

Mobile Computing—An Enabler in International Financial Services / *N. Raghavendra Rao* 828

Chapter LX

Mobile Computing: An Australian Case Study / *Paul Hawking, Gina Reyes, and Stephen Paull* 839

Chapter LXI

Introducing Mobile Technology into an Australian City Council: Experience and Lessons Learned /
Joanne Marie Curry 848

Chapter LXII

Emerging Mobile Technology and Supply Chain Integration: Using RFID to Streamline the
Integrated Supply Chain / *Richard Schilhavy and A.F. Salam* 859

Chapter LXIII

Mobile Batch Tracking—A Breakthrough in Supply Chain Management / *Walter Hürster,*
Hartmut Feuchtmüller, and Thomas Fischer 870

Detailed Table of Contents

Foreword xxxv

Preface xxxvii

Section I
Location

Chapter I
Information Management in Mobile Environments a Using Location-Aware Intelligent Agent
System / *Amrish Vyas and Victoria Yoon* 1

Mobile computing leads us into a new and fascinating journey into location-based services (LBSs) that was not feasible with land-based Internet connectivity. This is the dynamic creation of service offerings based on a location. However, the location itself can change from time to time. Thus LBSs form the crux of what is specific to mobile technologies, as discussed in this extremely well-researched and well-written chapter by Vyas and Yoon. The authors have rightfully argued that the recent rise in the level of comfort and demand to access various types of information using mobile devices can be attributed to the advancements in location-based services, which in turn are engendering new passion in mobile services utilizing users’ location information. Such spatio-temporal information processing entails the need for a dynamic middleware that accurately identifies changing user location and attaches dependent content in real time without putting extra burden on users. This chapter succinctly describes the creation of a distributed infrastructure that is capable of supporting such scalable content dissemination. The Location-Aware Intelligent Agent System (LIA) offered by the authors is a conceptual framework in integration with Publish/Subscribe middleware to comprehensively address dynamic content dissemination (based on PUSH and PULL strategies) and related issues.

Chapter II
Location-Based Services: Opportunities and Challenges / *Ramaprasad Unni and Robert Harmon*..... 18

As already stated, location-based services are expected to play an integral role in the mobile commerce domain. This ability of mobility in creating dynamic and location-based services opens up opportunities for mobile network operators and service providers to add value and create additional revenue streams through a variety of personalized services based on location of individual wireless users. This chapter makes a crucial and substantial contribution to the strategic thinking in this area of location-based services. The issues and

challenges discussed in this chapter include ownership of networks and their use by network operators and third parties, privacy concerns of consumers, and the corresponding business models for these services. The major areas covered by this chapter include an overview of location-based wireless services and their related technologies, a critical examination of the LBS value chain, and the strategic implications of location-based services for network operators and service providers.

Chapter III

Location Services in Cellular Networks / <i>Israel Martin-Escalona and Francisco Barcelo</i>	35

This chapter discusses the primary features of location services in cellular networks and mechanisms to implement them. The authors start with an excellent overview of the most important location-based services, followed by the main location techniques (including their constraints and mechanisms to overcome them) that facilitate the provision of these services. The solutions proposed in this chapter have been used by local regulatory bodies in their official recommendations. Finally, this chapter also reviews the location architectures standardized for use in the main cellular networks and presents the concept of location middleware as a natural addition to these architectures.

Chapter IV

Ontologies for Location-Based Services / <i>Matthias Brantner, Sven Helmer, Carl-Christian Kanne, and Guido Moerkotte</i>	54

An expressive and comprehensive service description is vital when offering Web services. This is so because the discovery relies on the ability to match a user's need accurately to a service description. Ontologies are a flexible and powerful method to describe services. In this chapter the authors demonstrate how ontologies can be used to improve service discovery considerably in a mobile context by offering location-based information. This discussion on ontologies is followed by an example ontology, and the authors explain how to integrate such an ontology into existing technologies, thereby providing an effective way to describe location-based services.

Section II Health

Chapter V

Relevance of Mobile Computing in the Field of Medicine / <i>Henrique M. G. Martins and Matthew R. Jones</i>	69

The healthcare domain stands to gain immensely by the incorporation of mobile information and communication technologies (MICTs)—as is demonstrated succinctly by the authors in this chapter. Researched by a medical doctor, this case study-based chapter delves deeper into the use of different MICT devices by doctors in specific hospital settings. While some doctors easily adopt MICT devices and find them a helpful tool, others encounter problems with their usage and, as a result, stop using the devices. This chapter identifies five factors influencing the uptake of MICTs in clinical work practices and proposes a framework for analyzing their interactions with the aim of increasing its uptake in medicine.

Chapter VI

Mobility in Healthcare for Remote Intensive Care Unit Clinical Management / *Carolyn*

McGregor 83

Computing and IT support within intensive care units (ICUs) has traditionally focused on monitoring the patients and delivering corresponding alarms to care providers within a hospital setting. However, many intensive care unit admissions are via intra- and inter-healthcare facility transfer, requiring receiving care providers to have access to patient information prior to the patient's arrival. The author discusses the opportunities that exist for mobile gadgets, such as personal digital assistants (PDAs), to substantially increase the efficiency and effectiveness of processes surrounding healthcare in the ICUs. This chapter provides invaluable reading and discussion on transcending beyond the current use of mobile devices in hospitals, which is restricted to mere personal information management and static medical applications, and takes the readers into the deployment of mobile-enabled solutions with overall considerations including privacy, cost, security, and standards.

Chapter VII

The Evaluation of Wireless Devices Used by Staff at Westmead Hospital, Sydney /

Sandra Synthia Lazarus 96

This chapter reports on a study to research and evaluate the use of latest generation wireless devices—typically personal digital assistant (PDA) devices—by clinical staff at the large Westmead Hospital located in the west of Sydney, Australia. Currently, medical reports in this and other hospitals are primarily recorded on paper supported by personal computers at nursing stations. However, there is very little or no access to medical reports and decision-making tools for medical diagnosis at the patient's bedside—the precise location at which most medical decision making occurs. Delays in access to essential medical information can result in an increased time taken for accurate diagnosis and commencement of appropriate medical management of patients. This chapter discusses the application of handheld devices into more powerful processing tools connected to a centralized hospital data repository that can support medical applications.

Chapter VIII

Mobile Comprehensive Emergency System / *Samir El-Masri* 106

This chapter discusses the application of Mobile Web Services in the handling of emergency processes in the health sector. The proposed application implements a mobile system based on cellular phone networks in ambulances, where communication between a number of parties is critical in terms of time, efficiency, and errors. Furthermore, it equips doctors with mobile devices that enable them to get connected to the Internet and access the health record of a patient quickly. This chapter demonstrates the way in which the proposed Mobile Comprehensive Emergency System (MCES) application would work with both static and mobile servers. The implementation of this new system will enhance the current system communication and makes it more reliable, consistent, and quick, and would also free the human intervention otherwise needed to access information.

Chapter IX

Application of Mobile Technologies in Healthcare Diagnostics and Administration / *Ketan*

Vanjara 113

This chapter explores various advancements in mobile devices and related software applications that facilitate rapid diagnostics in healthcare. Furthermore, this chapter also provides an excellent discussion,

based on the author's experience as well as study, on the incorporation and usage of mobile devices in healthcare administration. Finally, the integration and networking of mobile devices is presented as the next major and substantial level of development that would lead to comprehensive usage of mobility in healthcare.

Section III Technical

Chapter X

Energy-Efficient Cache Invalidation in Wireless Mobile Environment / *R. C. Joshi, Manoj Misra, and Narottam Chand* 132

This chapter discusses the caching at a mobile client as a technique that can reduce the number of uplink requests, lighten server load, shorten query latency, and increase data availability. A cache invalidation strategy presented by these authors can ensure that the data item cached into a mobile client has the same value as on the server of origin. Traditional cache invalidation strategy makes use of periodic broadcasting of invalidation reports (IRs) by the server. However, this IR approach suffers from long query latency, larger tuning time, and poor utilization of bandwidth. Updated invalidation report (UIR) is a method that replaces a small fraction of the recent updates as deemed necessary—thereby reducing the query latency. To further improve upon the IR- and UIR-based strategies, researchers present a synchronous stateful cache maintenance technique called Update Report (UR). The UR strategy outperforms the IR and UIR strategies by reducing the query latency, minimizing the disconnection overheads, optimizing the use of wireless channels, and conserving the client energy. This highly researched chapter is a ‘must read’ for readers researching and experimenting with caching techniques at the mobile client end.

Chapter XI

Review of Wireless Technologies and Generations / *Raghunadh K. Bhattar, K. R. Ramakrishnan, K. S. Dasgupta, and V. S. Palsule* 142

This is an excellent review of the basics of wireless communication and the corresponding wireless generations. While communications technology has advanced very rapidly during the last century, so has the needs and expectations of the people. The market has managed to balance the above scenario by providing the effective solutions as and when these became available, through a series of technological innovations. Furthermore, to encourage adoption and advancement in wireless communication technology, standardization of technologies and processes is also required. The era during which such technologies and standards are popular is generally termed as *Generations*. This chapter discusses the fundamentals of mobile technologies in relation to this mobile generation. Such a discussion of the communication technology generations not only provides an understanding of the past history of these technologies, but also creates the basis for understanding their future.

Chapter XII

Orthogonal Complex Quadrature Phase Shift Keying (OCQPSK) Spreading for 3G W-CDMA Systems / *Shailendra Mishra and Nipur Singh* 158

This chapter discusses the variety of digital modulation techniques that are currently being used in wireless communication systems, as well as various alternatives and advancements to these techniques. When using the 3G (third-generation) spread-spectrum systems, such as W-CDMA (3GPP) and cdma2000 (3GPP2), the handset can transmit multiple channels at different amplitude levels. However, modulation schemes such as OQPSK or GMSK do not prevent zero-crossings for multiple channels and are no longer suitable. There is

a need for a modulation format or a spreading technique that can accommodate multiple channels at different power levels while producing signals with low peak-to-average power ratios. This is what the authors have proposed through their own OCQPSK (Orthogonal Complex Quadrature Phase Shift Keying) spreading technique for W-CDMA and cdma2000. Starting with the basic structure of the reverse link (uplink) for W-CDMA and cdma2000 with no scrambling, this chapter explains the transition through complex scrambling to OCQPSK. The chapter then describes the concept of complex scrambling and OCQPSK and how it works. Finally, this chapter describes how to measure modulation quality on the reverse link of 3G systems and a complete downlink physical layer model, showing various results of BER and BLER calculation and also various time scopes and power spectrums.

Chapter XIII

Code and Time Synchronization of the Cell Search Design Influence on W-CDMA System

Performance / *Shailendra Mishra* 173

This chapter discusses the large-scale statistics of an Improved Cell Search Design (CSD) using cyclic codes and compares it with the 3GPP Cell Search Design using comma free codes (3GPP-comma free CSD) in terms of acquisition time for different probabilities of false alarm rates that would achieve faster synchronization at lower hardware complexity. The authors also propose design improvements in stage 2 of the 3GPP-comma free CSD by using a Fast Hadamard Transformer (FHT). Furthermore, masking functions are used in stage 3 of both the improved CSD and the 3GPP-comma free CSD to reduce the number of scrambling code generators required, as described in Chapter 13. This results in a reduction in the ROM size required to store the initial phases of the scrambling code and generators in stage 3, resulting in faster synchronization between the MS and the BS and improved system performance. The research results presented here indicate that for a channel whose signal-to-noise ratio is degraded with additive white Gaussian noise (AWGN), the Improved CSD achieves faster synchronization with the base station and has lower hardware utilization when compared with the 3GPP-comma free CSD scheme under the same design constraints.

Chapter XIV

Turbo Equalizer: A Solution to 4G Mobiles / *Priyatamkumar, R. M. Banakar, and*

B. Shankaranand 184

This research-intensive chapter discusses the performance gains for communications systems resulting from turbo codes, turbo equalization, and decoding techniques. Turbo codes send digital data over channels that require equalization—that is, those that suffer from inter-symbol interference (ISI). Turbo equalizers have been shown to be successful in mitigating the effects of inter-symbol interference introduced by partial-response modems and by dispersive channels for code rates of $R > 1/2$. The performance of iterative equalization and decoding (IED) using an M-BCJR equalizer is analyzed in this chapter. Furthermore, the bit error rate (BER), frame error rate simulations, and extrinsic information transfer (EXIT) charts are used to study and compare the performances of M-BCJR and BCJR equalizers on precoded and non-precoded channels. The authors predict the BER performance of IED using the M-BCJR equalizer from EXIT charts, and discuss in detail the discrepancy between the observed and predicted performances.

Chapter XV

Administration of Wireless Software Components / *Franck Barbier and Fabien Romeo* 200

The need for software standardization of mobile and wireless devices is crucial for successful component-based software engineering. Components in mobile and wireless devices require administration functionality that—despite existing standards, protocols, techniques, and tools—effective and efficient management may come up against odd component forms resulting from non methodical design. The availability of version 2 of

the UML (Unified Modeling Language) and the prominence of executable modeling languages in the MDA/ MDE (model-driven architecture/model-driven engineering) open up opportunities for building manageable wireless software components as discussed in this chapter. This chapter also discusses a design method and library created on top of the Built-In Test (BIT) technology, and illustrates the concepts through a case study of home automation systems.

Chapter XVI

Mobile User Data Mining and Its Applications / *John Goh and David Taniar* 216

This chapter discusses the issues of mobile user data mining. Mobile user data mining is the collection of data collected from the activities generated from mobile users in order to analyze their behavior pattern to predict their future behaviors. The increasing adoption of mobile devices provides the ability for mobile user data mining to analyze data collected from mobile users. This can be used to determine the trends and patterns for decision-making purposes. It is applicable to marketing, banking, and retail industries. Finally, this chapter provides an insight to the underlying issues in mobile user data mining.

Chapter XVII

Dynamic Matching of Supply and Demand in an M-Commerce Services Marketplace: Using Intelligent Agents and Semantic Technology / *Thomas Leary, A. F. Salam, and Rahul Singh*.... 233

Mobile users desire customized bundles of services that need to be dynamically created from the service providers. However, services are unique; and since unused services do not generate revenues, they present a lost ‘economic rent’ for organizations that are not part of the network of service providers. As a result, they are not part of the customized bundle of services. The dynamic discovery of a bundle of individual services from such a network that meets the unique needs and constraints of the mobile user requires intelligent agent technology. Such agent technology would match personal needs of the user with the available services in a cost-efficient manner. This chapter provides a mechanism to create dynamic service bundles from ad-hoc user requirements using intelligent agents. The authors apply this technique to a mobile commerce environment and illustrate the composition of user-specific service “bundles” by intelligent agents that represent the interests of the m-commerce user. Such agent-based architectures provide users with customized solution “bundles” that reduce their cognitive burden, while improving the utilization of resources for organization that are part of the service provider network.

Chapter XVIII

Information Delivery for Mobile Business: Architecture for Accessing Large Documents through Mobile Devices / *Christopher C. Yang and Fu Lee Wang* 247

In this information-centric age, an organization needs to access the most up-to-date and accurate information for fast decision making. Mobile access to the Internet provides convenient and portable access to a huge information space. However, loading and visualizing large documents on mobile devices is impossible due to their natural shortcomings such as screen space and computing power. In this chapter, the author introduces the fractal summarization model, based on fractal theory, for document summarization on mobile devices. This model generates a brief skeleton summary at the first stage, and the details of the summary on different levels of the document are generated on demand from users. Such interactive summarization reduces the computation load, which is ideal for wireless access. On the other hand, the hierarchical display in fractal summarization is more suitable for navigation of a large document, and it is ideal for small area displays. The automatic summarization, together with the three-tier architecture and the information visualization, are potential solutions to the existing problems in information delivery to mobile devices for mobile business.

Chapter XIX

Database Queries in Mobile Environments / *N. Marsit, A. Hameurlain, Z. Mammeri, and F. Morvan* 267

The technological evolution of networks, together with the development of positioning systems, has contributed to the emergence of numerous location-based services. These services will be of major technical as well as economical interest in coming years. This aroused the interest of a great part of the scientific community which proposed to study these services with diverse requirements and constraints. One of the direct consequences in the database field is the appearance of new types of queries (mobile queries issued from mobile terminals and/or requesting information associated with moving objects such as vehicles). This chapter proposes a survey on mobile queries, with particular attention given to the location issue.

Section IV Network

Chapter XX

A Nuerl Network-Based Mobile Architecture for Mobile Agents / *Anand Kuppuswami* 285

Wide area network (WAN) offers advantages such as providing myriad services available on globally diversified computers with reasonably simple process. The ability to dynamically create networks offers the processing powers of various processors at our command. With the advent of protocols like SOAP and Web Services, the consumption of services are more organized. In spite of various advances in communication techniques, the consumption of services through mobile gadgets is still only at the research level. The major impedances in implementing such systems on a mobile network are the latency factor, abrupt disconnection in service, lower bandwidth, and minimal processing power. The mobile agent's paradigm proves to be an effective solution to various issues raised. It has received serious attention in the last decade, and several systems based on this paradigm have been proposed and built. All such systems have been designed for a static network, where the service providers and the requestors are connected to the server on a permanent basis. This chapter presents a new framework of managing the mobile environment and the participating nodes with active intelligent migration. The functioning of the mobile agents in such a scenario is also presented.

Chapter XXI

Load Balancing as a Key to Enable Different Services in Cellular Network / *Sipra Das(Bit), and Sulata Mitra* 297

This chapter develops the concept of load balancing that plays a key role in providing various advanced applications in the cellular mobile environment. Load balancing means the efficient distribution of channels among cells in accordance with their requirements to minimize call blocking. As the channels for these services are scarce, load balancing has emerged as a primary issue in today's scenario. Two different prominent schemes of load balancing are elaborated upon. This chapter is aimed at researchers and policymakers, making them aware of the different means of efficient load balancing, as well as underscoring the problem areas that need further vigorous research.

Chapter XXII

A Novel Fuzzy Scheduler for Mobile Ad Hoc Networks / <i>S. Shanmugavel and C. Gomathy</i>	308

As mobile computing gains popularity, the need for ad hoc routing also continues to grow. In mobile ad hoc networks, the mobility of nodes and the error prone nature of the wireless medium pose many challenges, including frequent route changes and packet losses. Such problems increase the packet delays and decrease the throughput. To meet with the dynamic queuing behavior of ad hoc networks, to provide quality of service and hence improve performance, a scheduler can be used. This chapter presents a novel fuzzy-based priority scheduler for mobile ad hoc networks to determine the priority of the packets. The performance of this scheduler is studied using GloMoSim and evaluated in terms of quantitative metrics such as packet delivery ratio, average end-to-end delay, and throughput.

Chapter XXIII

Co-Existence of WLAN and WPAN Communication Systems / <i>Khaled Shuaib and Mohamed Boulmalf</i>	322

Next-generation wireless systems will provide users with a broad range of services, providing wireless technologies without any major interoperability issues. The recent growth in demand and deployment of WLAN/WPAN for short-range connections has been driven by the need to create ubiquitous networks, where one can be connected anywhere at any time making many services and applications a click away. These short-range access networks currently exist almost everywhere—at home, the workplace, hotels, hospitals, and so forth. Wireless local area networking standard (Wi-Fi) and the WPAN standard (Bluetooth and Zigbee) products utilize the unlicensed 2.4 GHz ISM band. Due to the dependence of these technologies on the same band, potential for interference exists. This chapter will focus on the characterization of these technologies, discussing differences and similarities, a wide range of applications and deployments, and the study of the potential interferences between such technologies when deployed within the same working space.

Chapter XXIV

Mobility Support Resource Management for Mobile Networks / <i>Mohammad Mahfuzul Islam and Manzur Murshed</i>	332

Mobile businesses are increasingly demanding high-speed facilities of multimedia services and Internet access “anywhere” and “anytime.” Limited transmission resources (i.e., bandwidth) are the main obstacles to widespread use of mobility in business. Many mobile networks support advanced technologies, and mobile communications protocols have been developed to optimally utilize wireless resources. These policies support the heterogeneous access technologies for multimedia services in mobile networks. Many of these policies exploit the mobility information from the current and the neighboring cells to dynamically adjust the resource reservation, allocation, and call admission control policy to adapt quickly with the changing network traffics. Resource reservation is, however, necessary to support the migrating users from the neighboring cells. This chapter explains the key components of resource management mechanisms in mobile networks, including the fair distribution of resources among different users/clients involved in mobile business or use of wireless resources.

Section V Security

Chapter XXV

Secure Agent Roaming for Mobile Business / *Sheng-Wei Guan* 366

This chapter proposes a secure agent roaming scheme in the m-commerce agent framework. Intelligent agents are one solution to providing intelligence in m-commerce. However, merely having an agent that is intelligent is insufficient for m-commerce applications. There are certain tasks that are unrealistic for agents to perform locally, especially those that require huge amounts of information. Therefore, it is important to equip intelligent agents with roaming capability, as is discussed in this chapter.

Chapter XXVI

Tackling Counterfeiting with a Secure Online Track-and-Trace System / *Fred Claret-Tournier, Pouwan Lei, Chris Chatwin, and Rupert Young* 379

This chapter introduces a secure online track-and-trace system for tackling counterfeiting. According to the Counterfeiting Intelligence Bureau (CIB), part of the International Chamber of Commerce, 7% of all world trade is in counterfeit goods, and the counterfeit market is worth \$350 billion. Virtually every country in the world suffers from counterfeiting, which results in lost tax revenue, job losses, health and safety problems, and business losses. Furthermore, counterfeit goods do not only target famous brand names, but anything that will sell, such as bottled water. Counterfeiters are increasingly damaging businesses, and as such, businesses need to fight back against counterfeiting. Nowadays, there is an explosion of mobile wireless services accessible via mobile phones with a built-in camera. The mobile users can access the Internet at any time, from anywhere, using ubiquitous inexpensive computing. Mobile camera phones and other handheld devices are becoming indispensable. The aim of this chapter is to show how businesses can protect their products from counterfeiting by using a secure online track-and-trace system, which will allow their customers to authenticate the products in real time through a Web-enabled mobile camera phone. This will assist business and customers by confirming that the said product is genuine and not counterfeit, which can be accomplished at anytime and any location, and without any significant changes to the existing business operational systems.

Chapter XXVII

Developing a Theory of Portable Public Key Infrastructure (PORTABLEPKI) for Mobile Business Security / *Sashi Nand* 393

This chapter reports on the development of a theory to increase the security of mobile business and its application to Australian information systems. To increase the growth of PKI, a theory called PORTABLEPKI is developed for the security of the wireless network. Furthermore, this chapter also discusses a framework for testing PORTABLEPKI and future research directions.

Chapter XXVIII

Systems, Handheld Devices, and Payment Methods for Mobile Commerce / *Wen-Chen Hu, Tom Wiggen, and Hung-Jen Yang* 401

The emergence of wireless and mobile networks has evolved the domain of electronic commerce in to a new application and research area that we know as mobile commerce. However, applying mobile commerce to business applications is a challenging task since it involves a wide variety of disciplines and technologies. In order to make it easier to understand the application of mobile commerce, this chapter starts the discussion

with the basics of mobile commerce from a technical perspective, followed by a discussion on Net-enabled mobile handheld devices such as smart phones and PDAs, and finally, mobile payment methods, including macro-payment and micro-payment methods.

Section VI Strategy

Chapter XXIX

Strategic Perspectives in Mobile Banking: Technology, Value Creation, and Developing Factors / *Achraf Ayadi and Chantal Ammi* 420

The convergence of the Internet and mobile networks has created new opportunities and applications. Considering mobile business only as an extension of the traditional Web can lead to missing out on unique and differentiable qualities for new value-added opportunities. Mobile banking is considered as potentially one of the most value-added and important mobile services available. The chapter examines the technological changes in mobile networks and the innovative attributes of a mobile Internet. It advances the theoretical framework of innovation in services to develop a customer-centric analysis of an m-banking value proposition. The chapter goes on to discuss critical factors in the diffusion of m-banking, and explores reasons of failure and further prospects of success.

Chapter XXX

Mobile Commerce in Emerging Economies / *Amol Patel* 429

This chapter discusses the opportunities and challenges of mobile commerce in emerging economies. It analyzes the profound impact of a mobile device on the way products and services are bought and sold in developing nations. The chapter argues that many mobile applications can have a much larger impact on emerging economies than those of the developed world. The chapter is aimed at creating an understanding of the unique social, technological, and economic drivers that can help entrepreneurs and solution providers to build and deploy compelling and revolutionary mobile commerce applications in these emerging markets.

Chapter XXXI

M-Business: A Global Perspective / *Mahesh S. Raisinghani* 435

This chapter discusses the use of mobile, handheld computer devices that are connected wirelessly to a network for business and personal use across people, projects, tasks, or organizational units to infer a trend of general acceptance of m-business in the marketplace. The author describes the state of the mobile commerce industry from a worldwide perspective and the barriers to implementation of m-commerce, discusses the issues and challenges, and ends with conclusions and directions for future research.

Chapter XXXII

Mobile Strategy Roadmap / *Francesco Falcone and Marco Garito* 444

This chapter takes the reader through a step-by-step process of developing a mobile business initiative. Starting by describing the fixed as well as Mobile Internet environment, this chapter analyzes the characteristics of a wireless world and how to incorporate mobility in business. The discussion is supported by current examples of successful implementations around the world, made by big and unknown companies. Eventually a tool to design and deliver a wireless solution is provided with an eye on the business side, trying to make technology and business work together and speak the same language.

Chapter XXXIII

Relating Mobile Computing to Mobile Commerce / *Nina Godbole* 463

This chapter takes a look at mobile commerce riding on the wave of mobile computing applications. Mobile commerce, also known as m-commerce, is the new powerful paradigm for the digital economy. In view of that, this chapter examines issues relevant to the mobility of today's workforce. The meaning of mobility and its implications are explored in this chapter, along with the legal implications that arise in the pursuit of mobile commerce. Given its importance, this chapter also briefly delves into security issues related to m-commerce. Towards the end, a lighter side of m-commerce and mobile computing is provided, together with conclusions and future directions.

Chapter XXXIV

The Future of Mobile Technologies and Applications in China / *Xiao Chen and Wei Liu* 487

This chapter deals with the future of mobile technologies and applications in China. The effect of emerging technologies, especially mobile technologies, on the massive market of China cannot be ignored in the global context. This chapter gives the reader an insight into China's mobile telecommunication industry today. The authors first relate statistics about China's mobile business market, including user and device analysis that helps in providing an understanding of mobile business in China. This analysis is followed by a description of the major mobile technologies employed in China and a brief view of the Chinese market's status, followed by an insight into some newly rising industries which are potentially successful mobile sectors in China. Finally, a real-life example is examined—that of the M-Government Project in Gunagzhou, capital city of Guangdong Province.

Section VII Application

Chapter XXXV

Developing Smart Client Mobile Applications / *Jason Gan* 504

Applications with rich user interfaces and smart clients improve the user experience. As mobile enabling technologies advance in capability, affordability, and availability, users expect improved design of mobile devices that will leverage the advances and convergence in technology and the Internet to deliver richer applications and value-added m-services. They demand m-applications that facilitate communications, information retrieval, financial management, paying bills, trading, gambling, entertainment, and dating. The design and architecture of the next generation of mobile applications and browsers will be challenging, as developers must still consider the limitations of the small screen and input options, and the unreliable connectionless paradigm, and allow for backward compatibility with earlier protocols and formats. Mobile application developers must support various configurations and interface with a plethora of different mobile computing devices and platforms. Furthermore, designers must also address any environmental and/or health issues, and design a product that is socially acceptable and safe.

Chapter XXXVI

Ontology-Based Information Retrieval Under a Mobile Business Environment / *Sheng-Uei Guan* 509

The establishment of the OntoQuery system in the m-commerce agent framework investigates new methodologies for efficient query formation for product databases. At the same time, it also forms new

methodologies for effective information retrieval. The query formation approach implemented takes advantage of the tree pathway structure in ontology, as well as keywords, to form queries visually and efficiently. The proposed information retrieval system uses genetic algorithms and is computationally more effective than iterative methods such as relevance feedback. Synonyms are used to mutate earlier queries. Mutation is used together with query optimization techniques like query restructuring by logical terms and numerical constraints replacement. Also, the fitness function of the genetic algorithm is defined by three elements: number of documents retrieved, quality of documents, and correlation of queries. The number and quality of documents retrieved give the basic strength of a mutated query, while query correlation accounts for mutated query ambiguities.

Chapter XXXVII

Intelligent Product Brokering Services / *Sheng-Wei Guan* 527

One of the potential applications for an agent-based system has been in the area of m-commerce, and a lot of research has been done on making the system intelligent enough to personalize its services for the user. In most systems, user-supplied keywords are normally used to generate a profile of the user. In this chapter, the author proposes a design for an evolutionary ontology-based product-brokering agent for m-commerce applications. It uses an evaluation function to represent the user's preference instead of the usual keyword-based profile. By using genetic algorithms, the agent tries to track the user's preferences for a particular product by tuning some of the parameters inside this function. The author has developed a prototype in Java, and the results obtained from these experiments look promising.

Chapter XXXVIII

Understanding Mobile Web Services (MWS) and Their Role in Integrating Mobile Devices / *Samir El-Masri* 544

Web services (WSs) have become the industry standard tools for communication between applications running on different platforms and built using different programming languages. The benefits, including the simplicity of use, that Web Services provide to developers and users have ensured integration of Web Services architecture by almost all IT vendors in their applications. As expected, with the proliferation of mobile phones, PDAs, and other wireless devices, the same requirements of making applications talk across platforms has become necessary on mobile devices. This has led to Mobile Web Services (MWSs), which are based on Web Services and related technologies like XML, SOAP, and WSDL, and which provide the best choice to be used in the architecture for integration of Web Services in mobile devices. This chapter discusses WS and MWS in the context of integration architecture, together with their advantages and disadvantages in usage. Since MWSs are deployed using wireless technologies and protocols, they are also presented and explained in this chapter.

Chapter XXXIX

Push Multicasting to Wireless Devices Using Publish/Subscribe Model / *Jon Tong-Seng Quah and Chye-Huang Leow* 553

Push technology is a kind of technology that automates the information delivery process without requiring users to request the information they need. Wireless has experienced explosive growth in recent years, and "push" is expected to be the predominant wireless service delivery paradigm of the future. For example, one would expect a large number and a wide variety of services such as alerts and messages as well as promotional contents and even e-mails to be regularly delivered to consumers' mobile devices such as phones or PDAs. As argued in this chapter, "pushing" information to a wireless device is a unique challenge because of the problems of intermittent communication links and resources constraint on wireless devices, as well as

limited bandwidth. The authors in this chapter explore an efficient multicasting mechanism that “pushes” pre-specified information to groups of wireless devices with limited bandwidth and flaky connections. This chapter reports on the design and implementation of a prototype framework based on the concept of push technology to multicast information via wireless technology.

Section VIII Method

Chapter XL

A Methodology for M-Transformation of Small and Medium Enterprises (SMEs) and its Application in Practice Using CBEADS© / *Ioakim (Makis) Marmaridis* 563

The advent of mobile technologies in recent times, coupled with the ever-increasing pressure for prices to drop, has opened up a whole new world of opportunities for business via the new medium for small and medium enterprises (SMEs). In particular, SMEs that have already embraced technology in many areas of their business find the move to embracing mobile technologies as the next logical step. This can be called m-transformation, and it consists of three ingredients: ICT infrastructure, new business process adoption, and a methodology to successfully lead to m-transformation and its many benefits for SMEs. The SME landscape presents some unique challenges however when it comes to attempting m-transformation. These challenges affect in turn all three aspects of m-transformation, causing the need for a methodology that is flexible and extensible in order to meet and surpass those challenges. In this chapter the author presents this methodology that has been used to successfully m-transform SMEs and shows that, although challenging, leading SMEs to a successful m-transformation is very possible, given sufficient background knowledge and a suitable, robust methodology to use.

Chapter XLI

Business Process Mobility / *Harpreet Alag* 583

In recent years, re-engineering of business processes is driven with a 360-degree view encompassing the customers, employees, suppliers, and partners. With the advancements in mobile technologies, mobile applications are swiftly making their way in enterprise business (processes). This chapter focuses on the application of mobile technologies in enterprise-wide business processes. The chapter particularly focuses on the use of mobile technologies to redesign or streamline business processes, including customer relationship management and supply chain management processes. The author has also succinctly highlighted how the “mobile layer” fit into the enterprise business architecture, and its subsequent potential.

Chapter XLII

Evaluation of Mobile Technologies in the Context of Their Applications, Limitations, and Transformation / *Abbass Ghanbary* 602

Emerging mobile technologies have changed the way we conduct business. This is because communication, more than anything else, has become extremely significant in the context of today’s business. Organizations are looking for communication technologies and corresponding strategies to reach and serve their customers. And mobile technologies provide the ability to communicate independent of time and location. Therefore, understanding mobile technologies and the process of transitioning the organization to a mobile organization are crucial to the success of adopting mobility in business. Such a process provides a robust basis for the organization’s desire to reach a wide customer base. This chapter discusses the assessment of a business in the context of mobile technology, presents a brief history of mobile technology, and outlines an initial approach for transitioning an organization to a mobile organization.

Chapter XLIII

Policy-Based Mobile Computing / *S. Rajeev, S. N. Sivanandam, and K. V. Sreenaath* 613

Mobile computing now encompasses the growing area of broadcast radio in data communication. This becomes an important criterion in providing good quality service with rapidly increasing mobile users. Policy-based approaches are widely used for security, quality of service (QoS), virtual private network (VPN), and so forth. In this chapter the authors examine the potential areas in mobile computing where policy-based approaches can be successfully implemented to enhance data communication.

Section IX Customer

Chapter XLIV

Investigation of Consumer Behavior in Using Mobile Payment Services—A Case Study of Mobile Recreational Services in Taiwan / *Maria Ruey-Yuan Lee, Yi-chen Lan, and Hsiang-Ju Su* 630

The growing popularity of the mobile phone and the diverse functionality of mobile services have forced mobile service providers to enter into a highly competitive business arena. In digital life today, mobile phone services are not restricted merely to communicating with people, but more and more value-added services have emerged to amalgamate disparate industries/businesses and open up greater market opportunities. These disparate industries/businesses may include recreational and travel services, mobile learning services, mobile banking services, and many others. Nevertheless the service providers must understand the consumer behavior in value-added services in order to enhance their product design. The key objectives of this research are to investigate and analyze the relationships between consumer behavior, consumer personality, and lifestyle in adopting mobile recreational services, and provide recommendations to the service providers for increasing competitiveness—in the context of Taiwan.

Chapter XLV

Mobile CRM: Reaching, Acquiring, and Retaining Mobility Consumers / *Chean Lee* 643

This chapter provides an introduction of using Mobile CRM to reach, acquire, convert, and retain consumers. Firstly, a definition of the term CRM is provided, and the author also gives an insight on extending CRM to the wireless world. Having presented the benefits of mobile data services and their benefits to businesses in terms of customer relations and marketing, however, businesses still faced the challenges of delivering the promise to consumers. More importantly, the adoption of mobile services is still low in business and consumer segments. The author identifies content appropriateness, usability issues, personalization, willingness to pay, security, and privacy as major challenges for businesses, and then recommends businesses to start segmenting their mobile consumers into Mobile Tweens, Mobile Yuppro, and Senior Mobile users; acknowledging that understanding the demographics, social, and behavioral issues of these three consumer groups is an initial step in Mobile CRM; before finally recommending the use of viral marketing as a mechanism to market mobile services. This is followed by matching relevant services to consumers to create a positive usability experience and always build a critical mass, but develop customers one at a time. The implementation of Mobile CRM will be fully addressed in the second part of the chapter.

Chapter XLVI

Factors Influencing Segmentation and Demographics of Mobile-Customers / *Anne-Marie*

Ranft..... 655

This chapter addresses important factors for consideration when readying a mobile commerce business for global business, addressing both regional differentiation in demographics that influence classifications of customer segments, and differentiation in demographics within a region. Globally, not all customer segments have regular access to mobile commerce facilities, and even for those that do, other demographic factors can impede their potential as mobile-customers. When starting from an Anglo-centric perspective, it is vital to have awareness of global differences in culture, language, payment options, time zones, legal restrictions, infrastructures, product needs, and market growth that could either improve or inhibit mobile-customer uptake, and in the worst case, result in unexpected litigation.

Section X Social

Chapter XLVII

Mobile Camera Phones—Dealing with Privacy, Harassment, and Spying/Surveillance

Concerns / *Christopher Abood* 666

This chapter discusses the growing inappropriate use of mobile camera phones within our society. There are two areas of concern that are dealt with in this chapter. The first concern deals with individual privacy and the use of mobile camera phones as a tool of harassment. The second concern deals with organizations seeking to prevent industrial espionage and employee protection. This chapter outlines how these devices are being used to invade individuals' privacy, to harass individuals, and to infiltrate organizations. The author outlines strategies and recommendations that both government and manufacturers of mobile camera phones can implement to better protect individual privacy, and policies that organizations can implement to help protect them from industrial espionage.

Chapter XLVIII

Social Context for Mobile Computing Device Adoption and Diffusion: A Proposed Research

Model and Key Research Ideas / *Andrew P. Ciganeck and K. Ramamurthy* 675

This chapter explores how perceptions of the social context of an organization moderate the use of an innovative technology. This chapter proposes a research model that is strongly grounded in theory, and offers propositions that investigate adoption and diffusion of mobile computing devices for business-to-business (B2B) transactions. Mobile computing devices for B2B are treated as a technological innovation. The authors believe that such an extension of existing models by considering the social contextual factors is necessary and appropriate in light of the fact that various aspects of the social context have been generally cited to be important in the introduction of new technologies. In particular, a micro-level analysis of this phenomenon for the introduction of new technologies is not common. Since the technological innovation that is investigated is very much in its nascent stages, there may not as yet be a large body of users in a B2B context. Therefore, this provides a rich opportunity to conduct academic research.

Chapter XLIX

A Socio-Cultural Analysis of the Present and the Future of the M-Commerce Industry /

Ritanjan Das, Jia Jia Wang, and Pouwan Lei 694

With high optimism, the third-generation mobile communication technologies were launched and adopted by telecommunication giants in different parts of the globe. However, with an uncertain and turbulent social, economic, and political environment, and the downturn in the global economy, difficult conditions are pronounced for the initial promises of m-commerce technologies to be fully realized. The causes for this, determined so far, have been largely of a technical nature. This chapter shifts the focus of analysis from a pure technical approach to a socio-cultural one. The basic premise of this chapter is that cultural variations do play a very important part in shaping potential consumers' choices, beliefs, and attitudes about m-commerce services. The authors believe that to be an important way for the m-commerce industry to fulfill its potential.

Chapter L

The Mobile Network as a New Medium for Marketing Communications: A Case Study /

Heikki Karjaluo, Matti Leppäniemi, Jari Salo, Jaakko Sinisalo, and Feng Li 708

This chapter discusses the mobile network as a new medium for marketing communications. It illustrates that the mobile medium, defined as two-way communications via mobile handsets, can be utilized in a company's promotion mix by initiating and maintaining relationships. Firstly, by using the mobile medium, companies can attract new customers by organizing SMS (Short Message Service)-based competitions and lotteries. Secondly, the mobile medium can be used as a relationship-building tool, as companies can send information and discount coupons to existing customers' mobile devices or collect marketing research data. The authors explore these scenarios by presenting and analyzing a mobile marketing case from Finland. The chapter concludes by pondering different future avenues for the mobile medium in promotion mix.

Chapter LI

Overview and Understanding of Mobile Business in the Age of Communication / *Joseph*

Barjis 719

This chapter studies mobile business in its dynamic, historic, and evolving nature. The chapter offers discussions on the background of, need for, and concept of mobile businesses. Following the background review, the chapter discusses the current status of mobiles business and its model. In this part, some classification of mobile business is given, and most representative fields of mobile business are identified, followed by a discussion on the technical aspects of mobile business. Elements that make a business mobile, such as communication infrastructure and supporting networks, are also discussed.

Section XI

Case Study

Chapter LII

Successful Implementation of Emerging Communication Technologies in a Mobile-Intense

Organization: A Case Study of Sydney Airport / *Keyurkumar J. Patel* 728

Wireless technology is growing at a phenomenal rate. Of the many present challenges highlighted by the author, increased security is one of the main challenges for both developers and end users. This chapter presents this important security aspect of implementing a mobile solution in the context of Sydney International Airport. After tackling initial challenges and issues faced during the implementation of wireless

technology, this chapter demonstrates how security issues and wireless application were implemented at this mobile-intense airport organization. The decision to deploy and manage the wireless spectrum throughout the airport campus meant that the wireless LAN had to share the medium with public users, tenants, and aircraft communications on the same bandwidth. Therefore, this case study also demonstrates an invaluable approach to protect unintended users from breach of existing security policies adopted by their corporate network. Authentication and data privacy challenges, as well as complete WLAN connectivity for tenants, public, and corporate usage is presented in this case study.

Chapter LIII

The Next Big RFID Application: Correctly Steering Two Billion Bags a Year Through
Today’s Less-Than-Friendly Skies / *David C. Wyld*..... 740

This chapter examines the adoption of radio frequency identification (RFID) technology in the commercial aviation industry, focusing on the role of RFID systems for improved baggage handling and security. Based upon secondary and trade literature, the chapter provides a timely overview of developments with regard to the implementation of the technology in commercial aviation. RFID technology holds distinct advantages over the currently used bar-code system for baggage handling. The chapter focuses on two major contributions that RFID promises commercial aviation: (1) improved customer service through better operational efficiency in baggage handling, and (2) increased airport and airline security. Particular attention is given to the initiative of Delta Airlines, an industry leader in the testing and development of RFID systems for improved operations in baggage handling. This chapter provides an avenue for academicians and business professionals to be aware of developments with RFID technology in this area.

Chapter LIV

Identified Customer Requirements in Mobile Video Markets—A Pan-European Case /
Torsten Brodt..... 754

Due to a significant cost advantage, mobile multicasting technology bears the potential to achieve extensive diffusion of mobile rich media applications. As weak performance of previous mobile data services suggests, past developments have focused on technology and missed customer preferences. Mobile multicasting represents a radical innovation. Currently, little research on consumer behavior exists regarding such services. The chapter addresses this gap by presenting results of qualitative and quantitative field trials conducted in three countries. It provides a continuous customer integration approach that applies established methods of market research to the creation of mobile services. Means-end chain analysis reveals consumers’ cognitive reasoning and conjoint analysis drills down to the importance of service attributes. Desire for self-confidence and social integration are identified key motivators for consumption of mobile media. Services should aim for technological perfection, and deliver actual and entertaining content. Interestingly, consumers appreciate reduced but tailored contents, and price appears not to be a superseding criterion.

Chapter LV

Applying Mobility in the Workforce / *Bradley Johnstone and Khimji Vaghjiani* 765

Due to a significant cost advantage, mobile multicasting technology bears the potential to achieve extensive diffusion of mobile rich media applications. As weak performance of previous mobile data services suggests, past developments have focused on technology and missed customer preferences. Mobile multicasting represents a radical innovation. Currently, little research on consumer behavior exists regarding such services. The chapter addresses this gap by presenting results of qualitative and quantitative field trials conducted in three countries. It provides a continuous customer integration approach that applies established methods of market research to the creation of mobile services. Means-end chain analysis reveals consumers’

cognitive reasoning and conjoint analysis drills down to the importance of service attributes. Desire for self-confidence and social integration are identified key motivators for consumption of mobile media. Services should aim for technological perfection, and deliver actual and entertaining content. Interestingly, consumers appreciate reduced but tailored contents, and price appears not to be a superseding criterion.

Chapter LVI

Applying Mobile Technologies to Banking Business Processes / *Dinesh Arunatileka* 778

This chapter discusses the impact of mobile technologies on service delivery processes in a banking environment. Advances in mobile technologies have opened up numerous possibilities for businesses to expand their reach beyond the traditional Internet-based connectivity, and at the same time have created unique challenges. Security concerns, as well as hurdles of delivering mobile services “anywhere and anytime” using current mobile devices, with their limitations of bandwidth, screen size, and battery life, are examples of such challenges. Banks are typically affected by these advances as a major part of their business deals with providing services that can benefit immensely by adoption of mobile technologies. As an example case study, this chapter investigates some business processes of a leading Australian bank in the context of application of mobile technologies.

Chapter LVII

Mobile GIS—Challenges and Solutions / *Pramod Sharma and Devon Nugent* 793

This chapter focuses on Mobile GIS (MGIS), which uses wireless networks and small-screen mobile devices (such as PDAs and smartphones) to collect or deliver real-time, location-specific information and services. Such services can be divided into field and consumer (location-based services) GIS applications. The use of wireless networks and small-screen devices introduces a series of challenges not faced by desktop or wired Internet GIS applications. This chapter discusses the challenges faced by mobile GIS (e.g., small screen, bandwidth, positioning accuracy, interoperability, etc.) and the various means of overcoming these problems, including the rapid advances in relevant technologies. Despite the challenges, many efficient and effective Mobile GIS applications have been developed, offering a glimpse of the potential market.

Chapter LVIII

Mobile Technologies and Tourism / *Pramod Sharma and Devon Nugent* 810

This chapter examines the potential of mobile technologies for the tourism industry. Mobile technologies have the capacity to address not only the pre- and post-tour requirements of the tourist, but also to support the tourist on the move. It is this phase of the tourist activity upon which mobile technologies can be expected to have the greatest impact. The development of applications for the mobile tourist will allow for the creation of a new range of personalized, location- and time-specific, value-added services that were not previously possible. Before such applications can be widely deployed, however, some fundamental technical and business challenges need to be addressed. Despite these challenges, mobile technologies have the potential to revolutionize the tourist experience, delivering context-specific services to tourists on the move as discussed in this chapter.

Chapter LIX

Mobile Computing—An Enabler in International Financial Services / *N. Raghavendra Rao* 828

This chapter suggests creation of a model for the financial services sector of the international financial market through the components of telecommunication and information technologies. While telecommunication and information technologies have influenced activities related to business, convergence of these

technologies is the crucial enabler that makes cross-border commerce in the present globalization scenario a reality. This chapter explains the use of knowledge-based financial systems and the process of incorporating mobile computing into these financial systems. The chapter discusses business process related to the financial services sector, creation of a knowledge-based financial system, and incorporating access to the system with devices that can be used in a wireless communication environment.

Chapter LX

Mobile Computing: An Australian Case Study / *Paul Hawking, Gina Reyes, and Stephen Paull* 839

This chapter uses a case study approach to demonstrate how companies are adopting mobile technology in their business processes. The authors describe a company that has used mobile devices to distribute work orders to field staff, and allows them to input their travel and work times and material usage for processing by the company’s ERP system. The chapter further examines the benefits obtained and the issues associated with the introduction of the system, and attempts to classify it according to an existing model.

Chapter LXI

Introducing Mobile Technology into an Australian City Council: Experience and Lessons Learned / *Joanne Marie Curry* 848

This chapter makes an important contribution in discussing the use of mobility in providing local government services. In early 2002, a large local government agency, Penrith City Council (PCC), located on the western fringe of the Sydney metropolitan area, entered into a collaborative working relationship with the University of Western Sydney. Research and development work conducted under this arrangement led to some interesting experiences and resulting learning for the students, client, and academic staff. This chapter highlights the development projects involved in the evolution of the PCC Mobile Strategy and discusses several aspects of the learning experiences, including: release hype vs. the implementation realities of mobile technology, technological options for the introduction of mobility, user acceptance of new technologies, the management of client expectations, and local government standards and guidelines and their impact on development directions.

Chapter LXII

Emerging Mobile Technology and Supply Chain Integration: Using RFID to Streamline the Integrated Supply Chain / *Richard Schilhavy and A. F. Salam* 859

This chapter explores how a mobile tracking technology is able to further streamline the integrated supply chain. Previous technologies which have attempted to integrate suppliers, manufactures, distributors, and retailers have lacked the flexibility and efficiency necessary to justify the prohibiting costs. Radio frequency identification (RFID) technology, however, enables various organizations along the supply chain to share information regarding specific products and easily remotely manage internal inventory levels. These applications are only a sample of what RFID is able to accomplish for the integrated supply chain, and this chapter seeks to explore those applications.

Chapter LXIII

Mobile Batch Tracking—A Breakthrough in Supply Chain Management / *Walter Hürster, Hartmut Feuchtmüller, and Thomas Fischer* 870

This chapter reports on a system solution that has been developed by T-Systems’ Solution and Service Center Ulm/Germany, within the service offering portfolio “Embedded Functions.” Considering that an increasing number of goods will be “on the road” (on rails, on ship, in the air) for an appreciable percentage of the

lifecycle, there is an urgent need to bridge the information gap between the automated systems at the factory sites and the storage control systems at the destination sites. This chapter discusses how the system solution has provided a synergy effect of connecting mobile communication solutions with auto ID services in the context of online surveillance during transportation, providing both downstream batch tracking, as well as upstream traceability.

Dedication

Keshav Raja

Foreword

As a conference chair of the recently completed International Conference on Mobile Business in Sydney, Australia, I had the unique opportunity of organizing, inviting, reviewing, and listening to a wide range of excellent researchers and practitioners in the area of mobile business. My understanding of mobility was further enhanced as I realised the phenomenal amount of research and industrial experimentation that is occurring in the area of mobile technologies and their application to business. Today, mobility encompasses themes such as devices, networking, architecture, design, applications, usability, security and privacy, entertainment, and mobile learning, to name but a few. These themes are embedded in this excellent book, edited by Dr. Unhelkar. This work, focused on the application of mobility in business, promises to become essential reading for all mobile researchers and practitioners. Therefore, with great pleasure and honour, I introduce this book to you.

While this handbook itself is research oriented, the contributing authors of the chapters within this book come from a cross-section of research *and* industrial expertise—a sensible and practical combination for a book that deals with *mobile business*. Through these internationally contributed chapters, the reader is exposed to in-depth discussions of the aforementioned mobility themes. The wide coverage of topics and the variety of contributors to this handbook make it a seminal addition to the body of literature and knowledge in the area of mobile business.

Mobile business, as the term indicates, deals with issues and challenges related to incorporation of mobile technologies in business processes. While the prior experiences gained from adopting electronic commerce can be helpful to business, mobile business provides some fundamentally unique issues and challenges of its own. For example, the customers in the mobile space are constantly on the move, requiring intensive research and experimentation in the area of location-based services to enable their tracking. Dropping of mobile connections is a common challenge in mobile transactions that also necessitates discussion of the security aspects of mobile applications. The sociological challenges of typing a text (SMS) reply on a small screen device or preventing an unauthorised photograph being taken in a swimming pool through a hand-held device open up a plethora of legal and ethical issues that are covered appropriately in the later chapters of this handbook.

Overall, I find this handbook an extremely comprehensive book dealing with the exciting and fast-moving domain of mobile business. Jonathan Withers, CTO of iBurst, reminded us (during his keynote address at the aforementioned conference) of the 10:30 Rule—in the next 10 years, technology will advance at a rate equivalent to the last 30 years! Researchers and industry practitioners will need to keep up with this challenge. However, this speed is what makes research in this area exhilarating. Cisco, the renowned networking company, further foresees fixed mobile convergence—where multiple applications will be seamlessly accessible from any access technology—as an area of research and application as we aim to provide a common set of services giving a consistent user experience regardless of the device.

My own philosophy of research in the mobile domain is that industry practitioners and academics *must* collaborate to understand, document, and provide practical solutions to the challenges of adopting mobility. To this end, I am particularly pleased to see the appropriate inclusion of numerous industry case studies in

this book. Furthermore, each of the chapters is of excellent quality, containing up-to-date and relevant information. Each chapter has also been extensively referenced so that readers may follow the references to obtain an even deeper understanding of the issues concerning mobile business. Finally, the international aspect of this book is another great plus for a work of this kind. In today's shrinking world, it is fitting that we share our experiences globally—as has been achieved in this edited work.

I highly recommend this book to both researchers and practitioners in the industry as an invaluable desktop reference. This book will not only aid practitioners in what they are currently doing with mobile business, but will also open up numerous directions for further investigative research work.

*Dr. Elaine Lawrence
University of Technology, Sydney, Australia
ICMB 2005 Conference General Co-Chair
August 2005, Sydney*

Preface

Communication is the key!!

THE FOURTH WAVE

The search for extraterrestrial life is becoming more urgent—lest we humans end up with claustrophobia! Jet travel could not have shrunk physical distance at the speed with which the Internet has meshed the world of words, sounds, and pictures. Businesses of all types and sizes find a level playing field in the cyberspace and, barring a few troglodytes of no particular age group, the world is connected. Alvin Toffler was right when he asserted in his popular book, *Third Wave*¹, that not only are we all affected by change, but even the *rate* of change is *accelerating*. A quick look around you as you read chapters from this book—sitting, traveling, working, or sleeping—ratify Toffler’s thoughts. And if, indeed, the rate of change is increasing, we are already *beyond* the third wave of information. This *Fourth Wave*, which is right upon us, is the wave of communication! Everyone is connected to everyone else *independent of time and location*. And mobile technologies provide the underlying basis for this age of communication, ensuring that businesses and people are connected directly and personally irrespective of where they are and what time of day or night it is. Riding on the back of the traditional Internet, mobile networks ensure that information that was available through a physical computing gadget at a fixed location is now available anytime and anywhere. This has obviously resulted in the tremendous popularity of mobile applications in the business domain. Furthermore, the *infrastructure* related to mobile technologies is also unique and, in some sense, distinctly different from the physical Internet. In simple business terms, setting of a transmission tower for mobile communications is relatively easier than installing physical lines to provide connectivity. Even the use of a satellite for communication purposes is becoming easier and cheaper for large organizations and governments than setting up of the costly fiber optics or similar physical wires and networks. Thus, because of their ease of usage from an end user’s viewpoint, ease of setting up and usage of infrastructure, and rapidly dropping costs, mobile technologies are influencing global business like never before. Based on this phenomenal importance of mobile technologies, especially to mobile business, this book compiles contributions from a wide range of researchers and practitioners in their investigations and usage of mobile technologies in business.

PARADIGM SHIFT

Advances in mobile technology are reshaping the relationship between business and technology. There is a “paradigm shift”—it is now no longer just a matter of providing technological solutions for business problems. This is so because mobile technology is now becoming a *creative cause* for hitherto unknown new business models and business processes. This leads to complexities in adopting mobility in business, as both technology

and business forces need to be balanced against each other. A similar balance is also aimed in the compilation of chapters in this book that bring together the technical, methodological, and social dimensions of mobile business to the fore. For example, while some chapters discuss the strategic needs of a business as a reason to use mobile technologies, other chapters demonstrate how mobile technology itself is considered as a driver for new business challenges as well as solutions.

The paradigm shift in terms of mobile technology adoption has occurred because, in practical terms, mobile technologies, including mobile applications, gadgets, networks, and content management systems, together work as a catalyst for deep structural change in how organizations accomplish goals. Although such a change requires significant effort in terms of BPM (Business Process Modeling), still the effort in applying mobile technologies to business processes and organizational structures is worthwhile as it allows organizations to gain greater reach and leverage new kinds of service delivery and interaction, culminating in significant productivity gains. In fact, it is not just organizations that stand to benefit with the advances in mobile technologies; the potential ease with which individuals can interact with organizations globally through mobile technologies also creates tremendous opportunities for easing the quality of life of people and society. Thus mobile technologies are becoming a creative cause for a paradigm shift in the business world that requires thinking, understanding, and case study experiences, as have been gelled together in this treatise.

MISSION

The aforementioned discussion highlights the crucial need to bring together the thinking and practical experiences of practitioners and consultants together with researchers and academics. This crucial need has been satisfied in this book, making it a significant contribution to the literature on usage and application of mobile technologies and networks in global business. This book provides significant *strategic* input into ‘mobility’, aiming to bring together thoughts and practices in technical, methodological, and social dimensions of this fascinating technology. Stated more succinctly, the mission of this book is:

To make a substantial contribution to the literature on ‘mobility’ encompassing excellence in research and innovation as well as demonstrated application of mobile technologies to mobile business.

CORE CONTENTS

For the sake of comprehensibility as well as enabling readers to focus on their area of interest, this book is divided into sections as follows:

- **Location:** Deals with location-based services that form the crux of the mobile revolution as applied to businesses.
- **Health:** Focuses on the phenomenal potential for application of mobility in the health sector.
- **Technical:** Focuses on the core mobile technologies.
- **Network:** Discusses the research and application of wireless and mobile networks.
- **Security:** Deals with the security aspect of mobile technologies.
- **Strategic:** Focuses on the strategic planning and management aspect of mobile technologies in business.
- **Application:** Discusses the incorporation of mobility in software applications.
- **Method:** Revolves around methodologies and processes related to mobile technologies including discussions on mobile transition processes.
- **Customer:** Focuses on the end user/customer aspects of mobility in business.

- **Social:** Handles the research as well as thoughts dealing the socio-cultural aspects of the influence of mobile technologies in society.
- **Case Study:** Discusses the practical application of mobility in practical business scenarios.

AUDIENCE

The following are the major categories of readers for this book:

- **Programmers and architects** of mobile-enabled software systems will find the discussions on technologies, networks, and security directly applicable to their work.
- **Business process modelers** and **information architects** will find the chapters dealing with incorporation of mobile technologies in business processes quite relevant.
- **Methodologists** and **change managers** will be interested in the chapters that describe the transition processes from existing to mobile businesses.
- **Sociologists** and **legal experts** will find the discussions on cross-border socio-cultural issues in applications of mobile technologies and the resultant globalization of businesses a fascinating read.
- **Strategic management** may find some of the earlier strategic discussions in this book quite relevant in setting the strategic directions of their organizations—especially because these chapters have been contributed by practicing senior managers.
- **Researchers** and **academics** will find numerous hooks in the research-based chapters of this book in terms of identifying areas of research, as well as following research methods when dealing with “mobility.” Thus, the strong research focus of this book—especially the detailed and relevant references at the end of each contributed chapter, the research methodologies followed, and the discussions on research results (especially some excellent “action research”-based case studies)—make this book an ideal reference point for active researchers in this area.

ENDNOTE

- ¹ Agricultural (first), Industrial (second), and Information (third) wave

Bhuvan Unhelkar
www.unhelkar.com

Critiques

Readers are invited to submit criticisms of this work. It will be an honor to receive genuine criticisms and comments on the chapters and their organization in this edited book. I am more than convinced that your criticisms will not only enrich the knowledge and understanding of the contributory authors and myself, but will also add to the general wealth of knowledge available to the ICT and mobile community. Therefore, I give you, readers and critics, a sincere *thank you* in advance.

Acknowledgments

First and foremost, I acknowledge all the wonderful contributing authors to this book. They come from a wide range of geo-cultural backgrounds, and they have enriched this book by making equally wide and varied contributions emanating from both intense research and practical industrial experience. This variation is reflected in the contents and presentation of the chapters. One thing in common, though—and which exudes through the individual chapters—is the dedication and hard work of each and every author in this book. Therefore, to these authors, I express my sincere gratitude and thanks.

Furthermore, I would specifically like to thank the following individuals (*my PhD students specifically for their help and work): Dinesh Arunatileka*, Bhargav Bhatt, Yogesh Deshpande, Anant Dhume, Samir El-Masri, Abbass Ghanbary*, Darrell Jackson, Vijay Khandelwal, Anand Kuppuswami*, Yi-Chen Lan, Girish Mamdapur, Javed Matin, Makis (Ioakim) Marmaridis, San Murugesan, Chris Payne, Anand Pradhan, Prabhat Pradhan, Mahesh Raisinghani, Prince Soundararajan, Ketan Vanjara, and Houman Younessi.

My thanks to my lovely family: wife Asha, daughter Sonki Priyadarashani, and son Keshav Raja, as well as my extended family, Chinara and Girish Mamdapur. This book is dedicated to my son, Keshav Raja, because of his disappointment and boredom with normal mobile phones at the ripe old age of 9+. All kids of around that age are going to grow up using mobile technologies in a way that is utterly unimaginable and hence I “pre-pay” my respect to them!

Bhuvan Unhelkar
www.unhelkar.com

Section I

Location

Chapter I

Information Management in Mobile Environments Using a Location-Aware Intelligent Agent System

Amrish Vyas

University of Maryland, Baltimore County, USA

Victoria Yoon

University of Maryland, Baltimore County, USA

ABSTRACT

Recent rise in the level of comfort and demand to access various types of information using mobile devices can be attributed to the advancements in wireless as well as Internet technologies. This demand leads us to the new era of mobile computing. Location-based services (LBS) are engendering new passion in mobile services utilizing users' location information. Such spatio-temporal information processing entails the need for a dynamic middleware that accurately identifies changing user location and attaches dependent content in real-time without putting extra burden on users. Our work focuses on creating a distributed infrastructure suitable to support such scalable content dissemination. As a result this chapter offers a conceptual framework, location-aware intelligent agent system (LIA) in integration with publish/subscribe middleware to comprehensively address dynamic content dissemination and related issues. We discuss the operational form of our framework in terms of PUSH and PULL strategies.

INTRODUCTION

The plateau of the information superhighway keeps advancing amid the evolution of the

Internet and related technologies. At the same time, popularity of mobile devices and rapid advancements in wireless technologies are making it convenient for users to access vari-

ous types of information available on the Internet over the wireless networks using their mobile devices. Moreover, as the array of mobile devices keeps expanding, users expect to be able to use different devices for accessing such information, entailing development of a research area called *mobile computing*. Although, mobile devices lack in terms of processing power, memory capabilities, display, connections to the wireless networks, and so forth, the demand for accessing dynamic content using mobile devices has grown ever more pressing (Kaasinen, 2003). On the other hand, timely and accurate data dissemination to and from various mobile devices using wireless networks and supporting technologies continues to be a progressively taxing research challenge.

Out of many challenging research issues in the mobile computing domain, a relevant challenge is context-aware computing. The term *context* refers to an application's operating environment, which consists of device location, device identity, user activity, time, state of other relevant devices, and so forth. Our focus in this chapter is on *location* and *time*. Location and time have a special relationship with regard to the content: *historical* (past) user locations and related content, *current* (present) user location and related content, and *future* user locations and related content. These scenarios represent the content usage as a function of location and time, giving rise to location-aware computing. Location-aware computing allows applications to be aware of a user device's physical location at any point in time. Applications can exploit this information for customizing their functional behavior and presentation. Users as well as providers of various types of mobile services can also rip the benefits of having access to this location information in a mobile environment. However, users are continuously moving along with their mobile de-

vices, and hence, location information of the user and her/his device is temporal. Capturing invariably changing location information of mobile devices presents an intriguing challenge. The Federal Communications Commission's (FCC's) mandate that wireless carriers in the United States be able to determine the approximate location of mobile phones making emergency calls is a key enabler for development of techniques to capture such temporal information regarding the user; it also provides an incentive to the cellular service providers to adopt above-mentioned location-aware systems. Examples of application of such location-aware systems include:

- **E-Deals:** A motel sends a promotional electronic coupon to mobile users passing by who are potential customers. Not only that, the motel can send some additional information regarding nearby restaurants and nearby attractions with applicable discounts if they take advantage of the e-coupon.
- **E-Directory:** A yellow page service that gives details on nearby services; for example, a user can locate the closest gas stations to her/his driving location, along with gas prices. Some additional information such as deals available on rotation of tires, car batteries, and so forth at a nearby auto center can be passed on to the user while presenting the user required information.

In performing the above tasks, location-aware systems need to combine the functionality of location-detection technologies (e.g., Global Positioning System, GPS), wireless or cellular telephone technologies (e.g., code division multiple access, CDMA), and information technologies (e.g., the Internet) under the scope of mobile computing to lay foundations for perva-

sive (anywhere, anytime) environments and services. On one hand, such services have the potential to dramatically improve usability of mobile devices and applications that adapt the content and presentation of services to each individual user and her/his current context of use. On the other hand, devising such location-aware systems is a tremendously complex task. Designers of location-aware systems have to keep in mind not only the continuous movement of mobile devices-related location-detection techniques, as well as connection with location-dependent content. Additionally, user perceptions about information privacy and security pose substantial challenges. System designers need to safeguard users' privacy in terms of making their location visible to the system all the time. They also need to secure users' personal information traveling through the system. Until users feel satisfied with the initiatives in this regard, the visions of innovative and powerful location-aware applications and services cannot be realized on a public network. Another challenge for system designers is the need to easily customize the existing Web content to specific geographic locations.

Location-detection techniques are one part of overall location-aware systems. There are several possible options for determining location of a mobile client, each requiring a different set of infrastructure and resulting in a different accuracy level. A few examples are: time difference of arrival (TDOA), angle of arrival (AOA), location pattern matching (LPM), Bluetooth technology, and the Global Positioning System (GPS). Out of these, TDOA, AOA, and Bluetooth are used either indoors or in limited range. GPS has emerged, recently, as not only a cost-effective, but widely acceptable locating technique that is also compatible with most wireless networks as well as information technologies. However, currently GPS carries an inability to function efficiently indoors and in

urban areas. Whilst RADAR (Bahl & Padmanabhan, 2000) or Bluetooth type technologies perform better in indoor areas, they cannot perform outdoors. Hence, an effective location-detection technology has to be a combination of these technologies so that outdoor as well as indoor locations can be effectively detected.

In addition to location-detection technologies, such systems also consist of wireless communication technologies such as cellular telephone technologies as well as information management technologies. In this chapter, we focus on developing information management techniques enabled with location-detection techniques. Location-aware applications, by default, are scalable distributed systems. We advocate using a middleware as a base for these distributed systems, as it provides not only the platform and protocols for communication, but also management supports making such systems as efficient and transparent as possible. We propose a *location-aware intelligent agent (LIA) system* that integrates two already proven components: agent technology and Publish/Subscribe paradigm. Agents append intelligence to an already flexible and scalable Publish/Subscribe architecture.

Publish/Subscribe (referred to as Pub/Sub from hereon) middleware has acquired attention and approval of researchers and is becoming popular tool for data dissemination in mobile environments (Anceaume, Datta, Gradinariu, & Simon, 2002; Baldoni, Beraldi, Querzoni, & Virgillito, 2004; Chen, Chen, & Rao, 2003; Farooq, Parsons, & Majumdar, 2004; Fenkam, Kirda, Dustdar, Gall, & Reif, 2002; Jose, 2004). It can offer distinguished assistance in extending the advantages of service-based architectures to the development of location-based services. The limited growth in such service-based architecture is mostly based on a *pull model*. A pull model is a user-initiated model in

which a user sends a request for information to the system and gets a response in terms of location-aware service offerings or answers. However, with advances in wireless Internet technologies and increasing competitive pressure amongst the service providers, a *push model* or service-initiated model is shaping up. Under the scope of such a push model, service providers actively push location-aware information to the users via the communication network according to users' predefined interests or historical usage data. Pub/Sub middleware is compatible with Internet technologies; however, by its nature it is not able to detect the location of the user and then connect such location information with related content while saving such information for future repeat usage.

The second component of our middleware is the innovative wave of intelligent agent technology. Software agents are beginning to play a pivotal role in our lives. Until recently, most of the research in *agent technology domain* was focused on modeling and designing agent-based systems. Researchers have recently started showcasing the applications of agent technology that can revolutionize many real-life problems. However, not only end-users but even researchers have wondered what exactly these agents are. Etzioni and Weld (1994) expressed that following the Information Superhighway metaphor, an intelligent agent can be:

- A backseat driver who makes suggestions at every turn, or
- A taxi driver who drives you to your destination, or even
- A concierge whose knowledge and skills eliminate the need to personally approach the superhighway at all.

Similarly, there are several other interpretations of the "agents" as an entity. Various definitions offered by researchers that portray

the variety of interpretations of agents are represented in the Related Work section. Given this difference in interpretation of what agents are or can be, we believe that that agent technology can play an important role in a distributed computing resources domain. Especially in our case it can provide a rather more robust middleware for wireless data dissemination in conjunction with Pub/Sub middleware, more so, considering the distinctive characteristics of mobile devices and/or usage patterns.

Next, we outline the related research work done so far in the area of agent technology, Pub/Sub middleware, as well as their usage in mobile environments. Secondly, we discuss the architecture of a proposed location-aware intelligent agent system, its components, and their functions. Thirdly, we discuss application of LIA in terms of push and pull strategies to an exemplary mobile services scenario, and finally conclude our discussion, along with some future research directions and a list of references.

RELATED WORK

Dynamic streams of information such as auction/stock trackers, traffic/weather alerts, and so forth communicated using mobile-distributed computing resources rely on an up-to-date view of information that can change rapidly and unpredictably. Much of this content may even be location dependent. Dissemination of such dynamic and location-dependent information to mobile clients has been a research challenge that researchers have been intrigued by for some time now. Some notable research efforts below have created direction for future research in this domain.

The first and an important aspect of location-aware systems is location-sensing techniques. Ladd et al. (2002) designed a location sensing system based on robotics using a wire-

less Ethernet. They began the design of the system by determining the position inside a building from measured RF signal strengths of packets on an IEEE 802.11b wireless Ethernet network. Using such a system they have tried to achieve reliable localization using general methods from Robotics following the Bayesian approach to localization. The system, however, was adoptable only within indoor locations. Smith, Balakrishnan, Goraczko, and Priyantha (2004), on the other hand, investigated the problem of tracking a moving device, which is a focal-point issue of location-aware computing. This investigation took place within the context of active as well as passive mobile architecture. They developed a real-world test-bed (Cricket) to evaluate the performance of location detection in both architectures. However, the Cricket system works indoors only, and as discussed above for a mass scale acceptance of location-aware systems, it needs to be as effective in outdoor locations as well. In addition to these research efforts, there are several other systems developed that have paved the path for further research. The active badge system is a classic example of such systems.

It is also valuable to predict future locations that the user will travel to, in order to make location-aware systems more useful, as well as wanted. Karimi and Liu (2003) focused their attention on a predictive location model for location-aware services. They submitted that under the scope of location-aware systems, users' future locations carry far more weight than usually assumed. There are additional benefits for the users and service providers when future locations can be predicted. They devised a PLM model that will predict users' future locations so that information dependent on such locations can be transferred to a user in advance to help with planned decision making. Within the model they used historical trajectories and a probabilistic matrix related to

road-level granularity of data for coming up with a prediction of future locations.

Kaasinen (2003) studied the need for location-aware mobile services from the user's point of view. The author drew conclusions regarding key issues in location-aware mobile services related to user needs based on user interviews, and laboratory and field evaluations with users and experts. They presented user needs under five main themes: topical and comprehensive contents, smooth user interaction, personal and user-generated contents, seamless service entities, and privacy issues. All these themes contribute to improving the overall usability of mobile services, applications, and in turn devices. Based on their findings, they suggested some guidelines for location-aware systems. For example, the services should be easy to find, and it should be easy for users to access an overview of the available services as well as their coverage. Also, services should be easy to take into use, and use thereafter. They also pointed out a few important aspects such as information personalization and user-generated contents.

Personalization in location-aware services provides a boost to usability of those services by providing the user with the most essential information according to the hierarchy of their personal preferences. However, the author acknowledges that if the user preferences are different in different locations, configuring the system for all available locations becomes a major task for users as well as system designers. On the other hand, they advocate active participation of users in information creation, rather than being just passive information consumers. Users' opinions, ratings, or recommendations could enhance many services. User participation is the cornerstone of development of the World Wide Web and other Internet services like Weblogs (blogs). The reason such paradigms became popular is because, among other factors, ordinary users can provide infor-

mation to others. However, it is not only difficult for users to participate, but most times users lack the motivation to do so. Our system (LIA), on the other hand, helps save the Pub/Sub middleware, and the agent framework supports such ordinary user participation.

A location-aware application consists of terminal and server components (Jarvensivu, 2004). The terminal component of LIA is the location-detection as well as device communication part of an agent framework that resides on user (end-users as well as service providers) devices. The server component on the other hand resides on fixed network resources and consists of Pub/Sub as well as all other parts of the agent framework including the system-wide repository. A proper coordination amongst these components enables smooth performance of the location-aware system. Scalability of these location-aware systems also determines their performance. Mokbel, Aref, Hambrusch, and Prabhakar (2003) defined scalability of location-aware systems as the system's ability to provide real-time responses to a large number of continuous concurrent spatio-temporal queries coming from the users to the central system. Mokbel et al. (2003) divided the spatio-temporal queries into *snapshot queries* and *continuous queries*. Snapshot queries are queries that can be answered using data that is already collected and saved on fixed computing resources. Continuous queries on the other hand depend for response on data progressively accumulating into the fixed resources. They also propose a sharing mechanism for these queries among various entities of the system in order to achieve the optimum scalability of the system.

In addition to sensing a user's current and precise location, as well as predicting possible future locations, there are other challenges for researchers. Schilit et al. (2003) discussed a few of the current challenges in the location-aware systems domain: low-cost, highly conve-

nient position-sensing technology, making users comfortable with respect to their location privacy, and having existing Web content easily customized to geographic locations. They initiated the Place Lab initiative to make location-aware computing a reality on a mass scale. The Place Lab initiative is meant to bootstrap location-aware systems through low-cost positioning technology in conjunction with a broad community-building effort that will create the large collection of location-enhanced Web services needed to catalyze business models.

On the other hand, passing information to users on their mobile devices has challenges of its own. For example, if the users' mobile device is turned off or disconnected from the network, how will the information be delivered to the user? Having a middleware that provides management support to the overall system is a solution to such practical problems. ELVIN (Carzaniga, Roseblum, & Wolf, 1998) is a notification system with limited support for mobile users. This system addresses the issues of message queuing, but important issues such as message distribution and location management are not addressed. CEA (Fenkam et al., 2002) and JEDI (Cugola, Nitto, & Fuggetta, 2001) are Pub/Sub middleware systems for mobile communications. Both these systems also address the queuing problem, specifically in cases of disconnections. The work of Huang and Garcia-Molina (2001) provided the operational guidelines for extension of Pub/Sub systems to a mobile environment. They analyzed the adaptation of a centralized and distributed Pub/Sub architecture with mobility. The ideas presented in Caporuscio, Carzaniga, and Wolf (2003), Chen et al. (2003), Huang and Garcia-Molina (2001), and Podnar, Hauswirth, and Jazayeri (2002) have provided us with motivation for our research efforts.

As noted above, a middleware is needed to effectively combine the functionalities of location-sensing technologies as well as wireless

technologies. We use Pub/Sub middleware as a platform to share such information amongst various technologies we use in our system. Pub/Sub middleware brings information providers and consumers together on a single platform. Publishers publish the information to the Pub/Sub system to be delivered to pre-identified interested receivers in the form of events or messages. This communication exhibits the following characteristics:

- **Anonymous:** The parties exchanging information do not have to identify themselves, nor do they have to know each other in order to send/receive the information (loosely decoupled).
- **Asynchronous:** The sender and receiver do not have to be connected to the system at the same time in order to exchange information.
- **Multicasting:** One publisher can publish the same information to many subscribers, and one subscriber can receive information from many publishers.
- **Stateless:** The system does not persistently store messages for all subscribers, rather messages are sent directly to those who have subscribed.
- **Implicit:** Receivers of information are chosen implicitly based on their subscriptions and cannot be altered or controlled by the publishers.
- **Dynamic:** Publishers can publish the most recent dynamic data they have available, and the Pub/Sub system will pass on the data to the recipients.

Congruent with our thinking, many researchers have acknowledged the potentials of applying a Pub/Sub system to wireless platforms. A few highlighting studies in this area include Caporuscio et al. (2003), Cilia, Fiege, Haul, Zeidler, and Buchmann (2003), Chen et al. (2003), Farooq et al. (2004), Heimbigner (2000),

Huang and Garcia-Molina (2004), and Muhl (2004). Much past research in this domain has focused either on development of Pub/Sub infrastructure or performance-oriented aspects of it. Caporuscio et al. (2003) proposed implementation of Pub/Sub systems with distributed access points where clients can connect with the system. This research focuses on enhancing the operational effectiveness of Pub/Sub-based infrastructure in a wireless environment. Farooq et al. (2004) have strictly studied the performance-related aspects of Pub/Sub middleware in mobile wireless networks. They studied publisher throughput, subscriber throughput, message loss, and handoffs, as well as workload parameters such as bandwidth, message and connection load, message size, and so on. However, most of above research has also acknowledged the fact that a platform only for sharing information amongst the service providers and service consumers is not enough; some management support is required to enhance the overall effectiveness of the middleware. We suggest integration of Pub/Sub with agent technology and harvest the benefits of both techniques to address such needs. Podnar et al. (2002) presented a multi-layered architecture that offers efficient content dissemination service targeting mobile users that motivated our work. However, their model lacked applicability to mobile environment, specifically in terms of location management, content adaptation, and information security. In our case, these features are addressed by use of intelligent agents. Our approach also differs from that of Podnar et al. (2002) in terms of component responsibilities, framework organization, and management of dynamic information processed throughout the proposed framework.

We avow that specific characteristics of intelligent agent technology can offer notable benefits within the scope of providing real-time services to mobile clients. In the past literature

many researchers have attempted to define agents, but ultimately highlight the variety of interpretations that exist in defining and/or describing agents.

Nwana (1996) defines ‘agents’ as: “A component of software and/or hardware which is capable of acting exactly in order to accomplish tasks on behalf of its user.” This definition follows the traditional definition of agency, agent accomplishing an assigned task by and for its owner. Bradshaw, in his book (1998) *Software Agents*, defines the term ‘agents’ in a more blunt way: “An agent is that agent does.” This definition is more in line with the trend of naming agent entities based on their functions or usage. Maes (1994) defines:

Autonomous agents are computational systems that inhabit some complex dynamic environment, sense and act autonomously in this environment, and by doing so realize a set of goals or tasks for which they are designed.

For the purposes of our research, we define “agent” as: “A knowledge-based, self-learning software entity that autonomously accomplishes tasks on behalf of its beneficiary(s).” The definition of agents, however, encompasses few but not all agent characteristics described as follows:

- **Autonomous:** Autonomy on the part of an agent means that an agent is able to take initiative and exercise a non-trivial degree of control over its own actions.
- **Goal-Oriented:** Agents have specific goals assigned either by the users (explicitly or tacitly) or by the designers at the time of design. They relate every action to the overall goal and have an ability to modify user requests, ask specific questions pertaining to user requests if they are of a destructive nature, or do not coincide with the ultimate goal.

- **Collaborative:** Agents cooperate and collaborate with other entities in the electronic environment. These entities include humans, other agents, other entities, and/or programs in the environment.
- **Flexible:** Agents’ actions are not predetermined in many cases. They deal with a set of incomplete and unprocessed information that often keeps changing over a period of time. Ideal agents will be designed to not only accumulate and relate to past knowledge, but constantly gather new information and design a response mechanism accordingly.
- **Self-Starting/Proactive:** Unlike other software programs which are explicitly invoked by user actions or any other event in the environment, agents can be proactive and exhibit a goal-directed behavior on their own.

Personalized/Customized

An agent can be personalized to a specific user, task, or environment. It can go through a complex process of customization on its own or by the user to be assigned a specific set of tasks or a general goal or both.

- **Reactive:** This refers to agents’ ability to selectively sense and act. Agents sense the changes occurring in their environment (the physical electronic world) and act accordingly, keeping in mind its goal.
- **Adaptive:** Agents are able to learn from their own actions, users’ preferences, and various elements of the environment, and fine tune their actions accordingly. This is based on the learning mechanisms built into agent design.
- **Communicative:** Agents can engage in simple to complex communications with other agents, users, and other applications in order to accomplish personal or com-

mon goals. Communications with users are carried out using natural languages, whereas the application uses machine languages and other communication protocols.

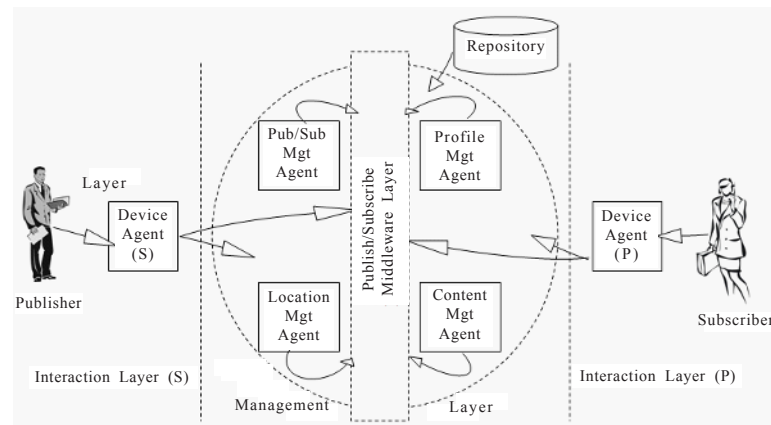
- **Mobility:** Agents can be built with mobility, transporting not only from one machine to another, but also across different systems, architectures, and platforms.
- **Inferential Capability:** Agents have models of self, users, and so forth based on which they exhibit a certain level of inferential capability. This means that agents not only have knowledge, but can infer upon that knowledge for taking specific actions.
- **Temporal Continuity/Long-Lived:** An agent is a continuously running program or process. As Etzioni and Weld (1994) describe, “It is not a ‘one-shot’ computation that maps a single input to a single output, then terminates.” In addition it follows some level of persistence in terms of its identity and state of being over a long period of time.

We argue that the above discussed qualities of intelligent agent systems in integration with Pub/Sub middleware make our system more robust, intelligent, and location aware in a mobile environment. To this end, we propose the *location-aware intelligent agent system* in integration with Pub/Sub middleware for mobile environments.

SYSTEM ARCHITECTURE

Figure 1 depicts the proposed multi-layered architecture of LIA. Because of various limitations we provide a distributed architecture which is a surrogate type of agent platform that provides management support to all other entities in the system. Hence, only device agents on the subscriber or publisher side are mobile, while the rest of the agent framework resides on fixed network resources. We chose this type of agent platform because of memory and processing capability constraints of mobile devices. The major components of our framework are: *Publishers*, *Subscribers*, and *LIA*.

Figure 1. Architecture of LIA



Publishers

Publishers are service providers who have an interest in reaching the consumers of the content they will publish in the form of events or messages. When a publisher is ready to deliver the content, it does so via Pub/Sub middleware. Assuming that the content matches the subscriber's interests as well as device specifications, the publisher defines the message/event to be published and transfers the content to the device agent-publisher (PA), who in turn sends a publish request to Pub/Sub. The publisher does not usually hold references to the subscribers, neither do they know how many of the subscribers the content will reach. We will learn below, however, that the PA keeps a log of published content as well as content requests that came from subscribers, keeping their decoupling intact. Publishers do not have to be connected to the system all the time or even at the same time when subscribers are connected to the system. Publishers could be host sites on a fixed network or can be mobile clients themselves. The message from one publisher can reach more than one subscriber, and also one advantage of our system is that a publisher does not have to keep publishing the same content over and over again.

Subscribers

Subscribers are content consumers. They are interested in receiving up-to-date dynamic information about their subjects of interest. Subscribers register their interests with the profile management agent (PMA) via device agent-subscriber (SA). The PMA shares these specific interests of the subscribers with the Pub/Sub middleware. This subscription information is not passed on to the publisher so that decoupling could be maintained. In addition to this, subscribers are also supposed to register with the PMA a number of devices they will be

using for receiving services such as a cell phone, PDA, notebook, or desktop computer. This will help the management layer locate the user device and recommend an appropriate list of location-aware services. Subscribers mainly use the Internet to connect, disconnect, and then reconnect to the system through various different access points based on the device they are using. It is noteworthy here that most mobile users will be faced with frequent disconnections and thus present a complicated challenge for the system for efficient message queuing, as well as delivery procedures.

LIA

LIA is an agent-based framework that is location aware as well as intelligent. LIA is composed of four layers: *Interaction Layer (P)* which is the interaction layer between LIA and the publishers, *Middleware Layer* which is the Pub/Sub middleware, *Management Layer* which provides the management support to the entire framework and is composed of four different types of agents, one system-wide repository, and finally *Interaction Layer (S)* which is the interaction layer between the subscribers and LIA.

Interaction Layer (P)

This interaction layer contains device agent-publisher (P). *Device Agent-P (PA)* resides on fixed network resources as long as the publisher is located in a fixed network as well. PA mainly communicates with the publishers and assists them in publishing their content to the service/information consumers. PA also assists publishers to dynamically offer location-aware dynamic information. The intelligent part of this service is how it helps the publishers dynamically configure and reconfigure the list of services in accordance with subscriber preferences. Not only that, based on subscription

requests from the subscribers, PA might suggest that the publisher offer a specific service in the close vicinity of physical or logical location.

Publish/Subscribe Middleware

Pub/Sub provides an interaction platform for publishers and subscribers. Publishers and subscribers can connect to the Pub/Sub middleware through their device agents (PA, SA). Pub/Sub middleware has a distributed structure that allows it to be scalable. Pub/Sub gets input from publishers, subscribers, and management layer. Publishers publish their content proactively or upon the request of subscribers to the middleware via PA. As such, Pub/Sub does not play an active role in adapting the content to users' interests; neither does it save any information related to the subscribers. However, Pub/Sub makes sure that only relevant information is published to subscribers. Pub/Sub enables multicasting by directing the same content to more than one subscriber. Pub/Sub middleware supports multicasting, asynchronous-anonymous communication, as well as time, space, and synchronization decoupling.

Management Layer

This layer provides management support to the overall architecture and is situated on fixed computing resources making all of its components stationary. It is composed of four intelligent agents with a distinct set of responsibilities—(1) *P/S Management Agent (PSMA)*, (2) *Location Management Agent (LMA)*, (3) *Profile Management Agent (PMA)*, and (4) *Content Management Agent (CMA)*—and a system-wide repository.

P/S Management Agent

PSMA is responsible for managing Pub/Sub operations, specifically, subscriptions and mes-

sages. It helps identify publishers/subscribers by locating identifier tags in either type of communication. A highlighting role PSMA plays is intelligent queuing of subscriptions or messages in a flexible way. In case of mobile device disconnection from the network, PSMA queues the subscription or message with time stamp and identifier information until reconnection is established. Also, when the content is delivered, it purges the queue. It requests and keeps a copy of all subscriptions or messages from the originating party to address message loss. It also generates knowledge by accessing publisher expertise as well as subscriber interests from the repository and makes proactive suggestions to both parties for better matching mechanism.

Location Management Agent

This component of agent framework helps locate publishers or subscribers who move along the network and connect through various access points. LMA is responsible for identifying all active devices registered to single user. Not only that, it is supposed to identify the user's geographical location while supporting several connect spaces (IP address, telephone numbers, DNS entries, etc.). There are many location-sensing technologies available: GPS, e911, Bluetooth, Active Badges, Cricket, and so forth. Although no single technology is likely to become dominant, as there are too many dimensions along which location-sensing mechanisms can vary. For its universal acceptance we use a GPS system. LMA works as a location server, and both SA and PA are receivers of geospatial information about the users with an inbuilt GPS receiver on their devices. These GPS receivers keep sending geospatial information for the user at every x minutes. Changes in users' direction, speed of traveling (assuming its constant), and all other pertinent information is recorded based upon this information.

As the GPS system transmits data in terms of coordinate values, a trajectory of traveling path is created based upon which information resources are organized. Once the momentary location and traveling path are identified, all information sharing is focused within the scope of such geospatial data. However, end-to-end control for such geospatial data is provided by the management layer considering the existent privacy and security concerns over users' information. The management layer filters already published information and seeks new information for the current as well as predicted future location (PFL) from repository as well as publishers. In addition to this, locations that users have visited in the past are also saved in the repository in case of any revisits. Supporting efficient rendering and transmission of geospatial representation will require attention to interaction issues associated with database access and knowledge discovery, which is supported by the management layer.

Profile Management Agent

PMA is responsible for organizing and communicating details regarding subscribers and publishers. PMA saves subscribers' interests, publishers' expertise, subscribers' registered devices, recent locations subscribers have been to, and so forth. In other words, every detail regarding publisher or subscriber is channeled through PMA to various components of the LIA. Both parties can choose a device (from a list of registered devices) and time for receiving/publishing the content, and register these interests with PMA.

Content Management Agent

CMA keeps a log of all messages and subscriptions that travel through the system. CMA deals with customizing the content as required. One of the most important responsibilities CMA carries out is to define associative rules for

related or complementary information to be presented in a way that is preferred by the user. CMA keeps record of each transmission, and when request for similar information comes from any other subscriber, CMA pushes the information to PSMA who in turn passes it on to the subscriber. CMA enforces content adaptation with respect to the active user device following its specifications. CMA also enforces information security using public-key protocol with all other components of the infrastructure. We assume that certification authorities can be built that can provide public key certificates prior to any communication between any components of the framework. We will also deploy a threshold secret sharing scheme offered in Shlolz (2002), where an (m,n) threshold scheme permits a message to be projected onto n shares such that any m of them can be combined to reconstruct the original message or subscription, but less than m of them can not. This will maintain the message security.

Interaction Layer (S)

This layer consists of device agent-subscriber (SA). Device agents are designed to interact directly with the end-users. SA resides mainly on the devices of the users; in this sense they are stationary agents. However, users also have them installed on their mobile devices; in that context they are mobile agents. SA communicates with users as well as the rest of the entities of the framework. Subscribers communicate their preferences and interests to the PMA through SA. SA keeps a log of all the services received by the subscribers on various devices. SA carries its "State" completely and folds into a mobile code when the user switches the device of active use and transfers itself onto the current active device. SA dynamically reconfigures the list of location-aware services in association with other agents in the system

and suggests any location-aware services users might be interested in while passing by that specific physical location.

Both publishers and subscribers can register more than one mobile device, and LIA carries its “State” completely and folds into a mobile code (transferred back to central architecture) when the users switch the device of active use and transfer itself onto the current active device.

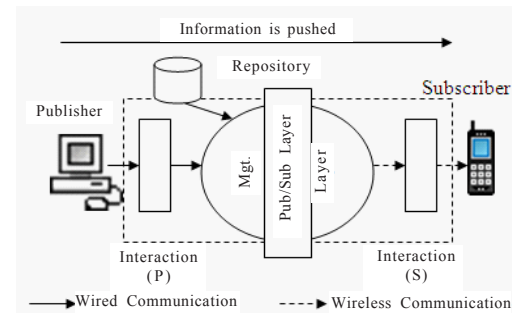
APPLICATION STRATEGIES

We outline the framework performance in view of an application scenario: Manni is driving on route 81-South near Salem, Virginia. She has her cat (Missy) with her. She is tired and wondering about availability of a specific type of room in a relatively inexpensive motel in the surrounding area. Fortunately, she is carrying her PDA with pre-installed device agent-subscriber (SA) as well as a location-sensing GPA receiver, which are components of LIA. SA is her gateway to LIA. We discuss two major strategies/models (Push, Pull) in light of this application scenario for location-aware dynamic data dissemination.

Push Strategy

Manni’s PDA is turned on and is online. Quality Inn, located in Salem, Virginia, is participating as a publisher with LIA. PA on QI’s side senses existence of Manni’s device in the range and pushes offers from QI along with types of rooms available to Pub/Sub middleware. Pub/Sub shares this information with the management layer. LMA identifies the precise location of Manni on I-81. PMA determines the fact that Manni will be using her PDA to receive her services today. CMA receives this input from LMA, PMA and processes the

Figure 2. PUSH strategy



content to match the location and adapt to fit the specifications of Manni’s device. PSMA identifies the message with its identifier and queues the final message for Manni’s device. Pub/Sub contacts the SA at this point and intimates that there is a message in queue waiting for Manni to read. Along with this information the management layer generates a list of additional services available for Manni. For example, CMA collects that Denny’s restaurant nearby offers early-bird discounts for Breakfast before 9 a.m., a golf club nearby is currently offering huge discounts to leisure guests, an area-wide directory service provides information on sightseeing places, car mechanics, grocery stores, departmental stores, shopping malls, and so forth. After receiving up-to-date information about all of these services, CMA bundles this information with the original message in a user-friendly manner.

Pull Strategy

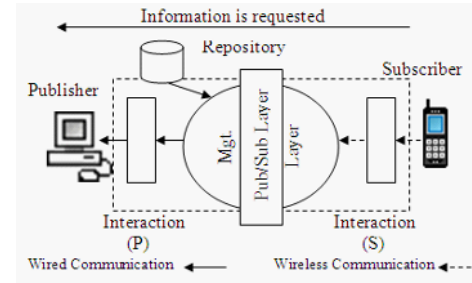
On the other hand, let us assume that Manni has forgotten Missy’s (the cat’s) food. Manni wants to find out if there are any specialty pet stores in the Salem area. Manni communicates this requirement and additional preferences to SA who shares the query with Pub/Sub. CMA accesses this information and explores the re-

pository to find out if there is already existing information in this regard. Assuming that there is no information available already, the management layer intimates the Pub/Sub to post a message to relevant publishers requesting such information. In addition, the management layer recognizes, with the help of PMA, that Manni likes to eat at Friendly's and has registered this interest with LIA. Pub/Sub locates service providers for the relative content and intimates them with the help of PA for generation of appropriate content for Manni. The service provider provides the content to Pub/Sub, who shares the information with the management layer. Through the same channel, information is communicated to Manni. By this time, let us assume that Manni has checked into a room at the motel and communicates her preference to receive the content on her laptop. PMA identifies her device, and CMA helps configure the content in a publishable form. LMA locates and identifies the notebook computer; CMA personalizes and reformats the information for an appropriate display on Manni's notebook computer. Figure 3 depicts the application of this strategy.

CONCLUSION

Dynamic content dissemination, particularly to mobile clients, is gaining an increasing amount of popularity. We have provided a framework which exhibits that agent technology, in association with Pub/Sub architecture, can make the system autonomous, intelligent, mobile, and secure, while keeping the benefits of the Pub/Sub paradigm. These are highlighting additions that agent technology can make to the research agenda at hand which are not documented to be delivered by other competitive technologies. Ours is the first effort to amalgamate contributions of agents and Pub/Sub in one system for effective performance in the mobile services

Figure 3. PULL strategy



domain. The synergetic effects of these two paradigms are also unprecedented.

Pub/Sub is a well-accepted solution for asynchronous and anonymous communications in a mobile environment. However, past literature left issues like system scalability, content adaptation, location awareness, and generation of dynamic content unanswered. Our framework, LIA, not only addresses all of the above issues, but extends the overall architecture to make it more robust. We believe that LIA will mark the future research direction for many intelligent and location-aware applications in the mobile services domain.

On the other hand, vulnerability of intelligent agents, especially in mobile environments, has been well documented. Although we have tried to address information security issues, we believe that it remains an open issue for future research. In addition, privacy issues relating to user-specific information requires further investigation.

In this chapter, we have chosen not to focus on the technical aspects and applicability issues of LIA due to the limitation of space and other resources. However, we believe that such discussion will be a driving factor in the acceptance of LIA, and we aim to encompass such discussion in our continuing research efforts. Another important limitation of our presentation so far is that we assume that the infrastructure required for successful operation of LIA is

not only existent, but at par with the performance measures. Integration of such infrastructure, including all three components—publishers, subscribers, and LIA—will mark the direction for ongoing research in this domain.

REFERENCES

- Anceaume, E., Datta, A., Gradinariu, M., & Simon, G. (2002). Publish/Subscribe scheme for mobile networks. *Proceedings of the Workshop on Principles of Mobile Computing*, Toulouse, France (pp. 74-80).
- Bahl, P., & Padmanabhan, V. (2000). RADAR: An in-building RF-based user location and tracking system. *Proceedings of the 19th Annual Joint Conference on the IEEE Computer and Communications Societies (INFOCOM)*, Tel Aviv, Israel (Vol. 2, pp. 775-784).
- Baldoni, R., Beraldi, R., Querzoni, L., & Virgillito, A. (2004). Subscription-driven self-organization in content-based Publish/Subscribe. *Proceedings of the International Conference on Autonomic Computing*, New York (pp. 332-333).
- Bradshaw, J. (Ed.). (1998). *Software agents*. Boston: The MIT Press.
- Caporuscio, M., Carzaniga, A., & Wolf, A. (2003). Design and evaluation of a support service for mobile, wireless Publish/Subscribe applications. *IEEE Transactions on Software Engineering*, 29(12), 1059-1071.
- Carzaniga, A., Roseblum, D., & Wolf, A. (1998). *Design of a scalable event notification service: Interface and architecture*. Technical report, Department of Computer Science, University of Colorado at Boulder.
- Chen, X., Chen, Y., & Rao, F. (2003). An efficient spatial Publish/Subscribe system for intelligent location-based services. *Proceedings of the 2nd International Workshop of Distributed Event-Based Systems*, San Diego, CA.
- Cilia, M., Fiege, L., Haul, C., Zeidler, A., & Buchmann, A. (2003). Looking into the past: Enhancing mobile Publish/Subscribe middleware. *Proceedings of the 2nd International Workshop on Distributed Event-Based Systems*, San Diego, CA.
- Cugola, G., Nitto, E. D., & Fuggetta, A. (2001). The JEDI event-based infrastructure and its application to the development of the OPSS WFMS. *IEEE Transaction on Software Engineering*, 27(9), 827-850.
- Etzioni, O., & Weld, D. (1994). A softbot-based interface to the Internet. *Communications of the ACM*, 37(7), 72-76.
- Farooq, U., Parsons, E., & Majumdar, S. (2004). Performance of Publish/Subscribe middleware in mobile wireless networks. *Proceedings of the 4th International Workshop on Software & Performance*, Redwood City (pp. 278-289).
- Fenkam, P., Kirda, E., Dustdar, S., Gall, H., & Reif, G. (2002). Evaluation of a Publish/Subscribe system for collaborative and mobile working. *Proceedings of the 11th IEEE International Workshops Enabling Technologies: Infrastructure for Collaborative Enterprises*, Pittsburgh, PA.
- Ge, Z., Ji, P., Kurose, J., & Towsley, D. (2003). Matchmaker: Signaling for dynamic Publish/Subscribe applications. *Proceedings of the 11th IEEE International Conference on Network Protocols*, Atlanta, GA (pp. 4-7).
- Heimbigner, D., (2000). *Adapting Publish/Subscribe middleware to achieve Gnutella-like functionality*. Technical Report, Department of Computer Science, University of Colorado at Boulder, USA.

- Huang, Y., & Garcia-Molina, H. (2001). Publish/Subscribe in a mobile environment. *Proceedings of the International Workshop on Data Engineering for Wireless and Mobile Access*, San Diego (pp. 27-34).
- Jarvensivu, R., Pitkanen, R., & Mikkonen, T. (2004). Object-oriented middleware for location-aware systems. *Proceedings of the 19th Annual ACM Symposium on Applied Computing*, Nicosia, Cyprus (pp. 1184-1190).
- José, R., Moreira, A., & Rodrigues, H. (2003). The AROUND architecture for dynamic location-based services. *Mobile Networks and Applications*, 8, 377-387.
- Kaasinen, E. (2003). User needs for location-aware mobile services. *Pers Ubiquit Comput*, 7, 70-79.
- Karimi, H., & Liu, X. (2003). A predictive location model for location-based services. *Proceedings of the 11th ACM International Symposium on Advances in Geographic Information Systems*, New Orleans, LA (pp. 126-133).
- Ladd, A., Bekris, K., Rudys, A., Marceau, G., Kavraki, L., & Wallach, D. (2002). Robotics-based location sensing using wireless Ethernet. *Proceedings of the 8th Annual International Conference on Mobile Computing and Networking*, Atlanta, GA (pp. 227-238).
- Maes, P. (1994). Agents that reduce work and information overload. *Communications of the ACM*, 37(7), 31-40.
- Martin-Flatin, J. P. (1999). Push vs. pull in Web-based network management. *Proceedings of the 6th IFIP/IEEE International Symposium on Integrated Network Management (IM'99)*, Boston (pp. 3-18).
- Mokbel, M., Aref, W., Hambrusch, S., & Prabhakar, S. (2003). Towards scalable location-aware services: Requirements and research issues. *Proceedings of the 11th ACM International Symposium on Advances in Geographic Information Systems*, New Orleans, LA (pp. 110-117).
- Mühl, G., Ulbrich, A., Herrmann, K., & Weis, T. (2004). Disseminating information to mobile clients using Publish-Subscribe. *Data Dissemination on the Web, IEEE Internet Computing*, 8(3), 46-53.
- Nwana, H. (1996). Software agents: An overview. *Knowledge Engineering Review*, 11(3), 1-40.
- Padovitz, A., Zaslavsky, A., & Loke, S. (2003). Awareness and agility for autonomic distributed systems: Platform-independent Publish-Subscribe event-based communication for mobile agents. *Proceedings of the 14th International Workshop on Database and Expert Systems Applications*, Prague, Czech Republic (pp. 669-673).
- Podnar, I., Hauswirth, M., & Jazayeri, M. (2002). Mobile push: Delivering content to mobile users. *Proceedings of the 22nd International Conference on Distributed Computing Systems*, Vienna, Austria (pp. 563-370).
- Schilit, B., LaMarca, A., Borriello, G., Griswold, W., McDonald, D., Lazowska, E., Balachandran, A., Hong, J., & Iverson, V. (2003). Challenge: Ubiquitous location-aware computing and the "Place Lab" initiative. *Proceedings of the 1st ACM International Workshop on Wireless Mobile Applications and Services on WLAN Hotspots*, San Diego, CA (pp. 29-35).
- Schlolz, J., Grigg, M., Prekop, P., & Burnett, M. (2003). Development of the software infrastructure for a ubiquitous computing environment—the DSTO iRoom. *Proceedings of the*

Workshop on Wearable, Invisible, Context-Aware, Ambient, Pervasive and Ubiquitous Computing, Adelaide, Australia.

Smith, A., Balakrishnan, H., Goraczko, M., & Priyantha, N. (2004). Tracking moving devices with the Cricket location system. *Proceedings of the 2nd International Conference on Mobile Systems, Applications and Services*, Boston (pp. 190-238).

Sutton, P., Arkins, R., & Segall, B. (2001). Supporting disconnectedness—transparent information delivery for mobile and invisible computing. *Proceedings of the IEEE Interna-*

tional Symposium on Cluster Computing and the Grid, Brisbane, Australia (pp. 277-285).

Wood, M., & Glade, B. (1996). Information servers: A scalable communication paradigm for WAN and the information superhighway. *Proceedings of the 7th Workshop on Systems Support for Worldwide Applications*, New York (pp. 305-310).

Vyas, A., & Yoon, V. (2004). Location-aware intelligent agent system (LIA) for Publish/Subscribe middleware in mobile environments. *Proceedings of the 3rd Workshop on E-Business (Web '04)*, Washington, DC (pp. 75-81).

Chapter II

Location-Based Services: Opportunities and Challenges

Ramaprasad Unni

Portland State University, USA

Robert Harmon

Portland State University, USA

ABSTRACT

Location-based services are expected to play an integral role in the mobile-commerce domain. Mobile network operators and service providers have the opportunity to add value and create additional revenue streams through a variety of personalized services based on location of individual wireless users. However, strategic thinking in this area is still evolving. Many issues surrounding location data such as ownership and their use by network operators and third parties, privacy concerns of consumers, and business models for these services are not well understood. This chapter provides (1) an overview of location-based wireless services and their related technologies, (2) an examination of the LBS value chain, and (3) strategic implications for network operators and service providers.

INTRODUCTION

With the roll-out of 3G networks and related infrastructure, the hype surrounding location-based services (LBSs) is finally beginning to get translated into reality. These services represent a significant development in the mobile business world. These services use location

data of mobile customers to provide a variety of location-specific services. The data and shopping services that can be tailored to a specific consumer's location and time parameters clearly differentiate the mobile experience from the wired experience and open the possibility for creating a wide range of new applications. Unlike many existing wireless data applications

that are merely Internet based, location-based services are specifically based on ability to dynamically locate the user (Kalakota & Robinson, 2002).

The success of these services is critical for mobile operators for several reasons. First, many of these operators have made large investments running into several hundred million dollars in technology relating to determining locations of mobile users. It is logical for network operators to seek returns on their investment through commercialized location-based services. Second, LBSs provide a means to create competitive advantage in an increasingly competitive market. Finally, LBSs represent a viable route to create value and increase average revenue per user (ARPU) through new revenue streams. Current indications are that ARPUs might go up 4 to 5% per year (Faggion & Trocharis, 2004). However, introduction of these services has been slow.

One of the reasons for the slow growth is failure on the part of carriers to invest in resources beyond the basic infrastructure that can enable the full potential of LBSs to be realized. Strategic thinking in this area is evolving, and there is a need to understand implications to different players in the LBS value chain. Key issues such as the ownership and management of location-specific data, transaction and data security, and consumer privacy are still to be resolved.

There are three main goals of this chapter. First, it provides an overview of location-based mobile-wireless services and related technologies. Second, it provides an examination of the LBS value chain and the dynamics within this sector. Finally, it examines the strategic opportunities and challenges for key players in the LBS value chain. The primary focus is the emerging LBS sector in the U.S. market.

BACKGROUND

Spatial and geographic information have been used by businesses for store location and merchandizing decisions since the advent of modern retailing in the early twentieth century. Location and physical proximity were important factors for attracting customers and improving sales (Christensen & Tedlow, 2000). The rapid adoption of electronic commerce seemingly made such location information less important because consumers could shop anywhere in the world without stepping out of their homes. However, the new technologies of location-enablement make location information mission critical in ways that were simply not possible earlier. It enables businesses to potentially offer timely and highly personalized services that are location specific. For the first time, marketers would be able to link existing knowledge of the consumer's identity, financial status, and buying history with the LBS parameters of a purchase including its exact time, place, purchasing behavior, and situational context as it happened in real time.

Mobile location services have evolved rapidly over the last few years. The difference between the earlier services and the new services lies primarily in the generation of location data. The former requires the manual input of location data, such as street intersection or zip code. This approach does not require investment in special location equipment, and generates less concern over privacy issues. However, the weaknesses of such systems are apparent. Consumers may not know their location, and consumer-supplied location information limits the ability of marketers and network operators to proactively offer a range of personalized services (Robinson, 2000). For the new services, the location data is automatically generated and updated by the network or the

device. This feature enables specific services such as location-dependent advertising and promotion to be triggered when the mobile device is in a specific pre-defined area.

In the U.S., the FCC's E911 mandate that requires network operators to facilitate emergency services by providing automatically generated precise cell phone location data provided the impetus for new LBSs. The network operators have invested several hundred millions of dollars in building up the necessary infrastructure. Operators are seeking to commercialize this location-enablement technology in order to leverage their investment. The FCC does not require the mobile network operators to use a specific technology. However, it has indicated specific performance metrics for location-enablement technologies. For network-based

technology, location information accuracy in Phase II must be within 100 meters 67% of the time and within 300 meters 95% of the time. For handset-based technology, location information accuracy must be within 50 meters 67% of the time and 150 meters 95% of the time (Burnham, 2002). The FCC has extended the deadline for implementation of Phase II several times. The current deadline is December 31, 2005 (Angelides, 2005).

TECHNOLOGIES

Location-enablement technologies are network based, handset based, and hybrid in nature. Network-based technologies use the cellular network to determine the location of the mobile

Table 1. Location-enablement technologies

TECHNOLOGY	DESCRIPTION	ADVANTAGES	DISADVANTAGES
Network-Based			
Cell-of-origin (COO)	Information generated about the cell occupied by a user	RF technology Inexpensive—uses existing network No handset modification Fast implementation No consumer behavior change	Low resolution
Angle of Arrival (AOA)	Measures angle of signal from mobile device to cell towers, minimum of 2 cell sites required	RF technology No handset modification No consumer behavior change	Expensive network modifications required. Resistance toward more antennas in neighborhoods Line-of-sight constraint Medium resolution (not less than 150 m)
Time Distance of Arrival (TDOA)	Triangulate at least 3 stations to measure and compare arrival time of signal from a user	RF technology No handset modification No consumer behavior change	Line-of-sight constraint Expensive Medium resolution Appropriate for CDMA
Enhanced cell ID (E-CID)	Software-based solution that compares list of cell sites available to user and checks for overlaps	RF technology Line-of-sight not required Moderate cost to upgrade	Works only with GSM Some modification required in handset and network
Handset-Based			
Global Positioning System (GPS)	Radio-navigation system comprising 24 low orbit satellites, triangulation by measuring the time to communicate with three satellites	RF technology Very accurate, 1-5 meters, 95% precision Not dependent on network	Line-of-sight issues Significant handset Handset modification May require consumer behavior change modifications
HYBRID TECHNOLOGY			
Enhanced Observed Time Difference (E-OTD)	Similar to TDOA, but handset calculates the location	RF technology Accuracy of 50 to 125 m. Some behavior change	Suited for GSM only Network and handset modification Cell coverage necessary
Assisted Global Positioning System (A-GPS)	Processing done by network while using the satellites	RF technology Moderate modification to handset Line-of-sight constraint reduced	Significant changes to network

device. Handset-based technologies use the radio navigation system provided by the satellites of the Global Positioning System (GPS). Most network operators (carriers) are now evaluating hybrid technologies that use both the network and the GPS system (Akçayli, Brooks, Laszlo, & McAteer, 2001). Additionally, other technologies such as the Ethernet-based wireless WLAN 802.11 may have the potential to offer certain types of LBSs like location-based advertising and promotion. The technologies are briefly described in Table 1.

Network-Based Technologies

The location technologies are based on the parameters of transmission such as signal propagation time and angle of arrival. These technologies typically require considerable expenditure on the network infrastructure, but do not require any modifications on the handset.

Handset-Based Technologies

GPS-based technology is a device-centric technology that detects how far it is from at least three satellites of the U.S. government-operated Global Positioning System. It is accurate to within 10 to 20 meters. However, it has several drawbacks such as increased cost, size, and power consumption of the mobile device. It also requires line of sight to calculate location.

Hybrid Technologies

Network-assisted GPS hybrid technologies are expected to deliver the accuracy of GPS and overcome the drawbacks of GPS associated with its line-of-sight requirement, and power consumption by shifting significant processing load from the device to the network (Burnham, 2002).

Short-Range Technologies

In addition to the above technologies, WLAN (or Wi-Fi) and Bluetooth technology have the potential as platform enablers for location-based services, such as personalized mobile advertising and promotion. These technologies are limited in their coverage and as such cannot be viewed as competing with network-based or GPS technologies. However, they are good for coverage in small areas. Additionally, the growth in WLAN adoption and the emergence of aggregators like Boingo and entry of mobile operators in the Wi-Fi field may offer seamless connectivity and location-enablement over larger areas (Blackwell, 2001). Retail establishments such as Starbucks have already invested in setting up such Wi-Fi networks to allow Internet access to their customers. Mobile devices would require hardware and software additions to use the wireless network.

Bluetooth technology presents a standard for wireless voice and data transfer over short distances, usually 8 to 10 meters at up to 1 Mbits/s. It also operates in the unlicensed 2.4 GHz band. It does not need a direct line of sight. Bluetooth chipsets are included in many brands of PDAs, cell phones, and other mobile devices.

Bluetooth and WLAN technologies do not comply with the FCC mandate. However, they do offer the potential to identify and communicate with a customer within a radius of a business establishment. A flexible approach that uses complementary technologies is likely to be adopted to provide accurate information in a timely fashion (Angelides, 2005). In this chapter, the focus is primarily on location services that use technologies other than Bluetooth and WLAN.

Table 2. Consumer-oriented location-based services

LOCATION-BASED SERVICE	CONSUMER APPLICATION
Information/directory services	Dynamic yellow pages service that automatically informs customer of location of nearest parking facilities, restaurants, etc. Travel, show, dinner reservations. Concierge services.
Tracking services	Tracking of children by parents. Locating lost pets. Locating friends in a geographic area. Tracking stolen cars. Tracking assets.
Emergency services	Roadside assistance. Search-and-rescue missions. Police and fire response. Emergency medical-ambulance.
Navigation	Route description. Dynamic navigational guidance. Traffic status in the area.
Location-based advertising and promotions	Wireless coupon presentation, targeted ads, and promotional messages keyed to the location. Promotional alert when a sale of a desired product takes place. Customer identification in a store or a neighborhood.

TYPES OF LBSS

LBSs are broadly defined as services that are enhanced by and depend on information about a mobile device's position. Location information is used to filter out irrelevant information and provide the context for different services. These services could be offered and executed both within and outside the mobile operator's network (Paavilainen, 2002). For consumers, LBS applications include emergency and safety-related services, entertainment, navigation, directory and city guides, traffic updates, and location-specific advertising and promotion, in addition to site-based purchasing with e-wallet-enabled wireless devices. Examples of these services are provided in Table 2.

Operators and service providers in the U.S. market are also looking towards the business customer and movement of data as an important route for profits (Reid, 2004). Autodesk recently launched mobile resource manager, a location-enhanced service that businesses use

to locate, manage, and communicate with their mobile workforce. Using this service, managers can find out locations of different employees, and determine who would respond to a request from a customer site.

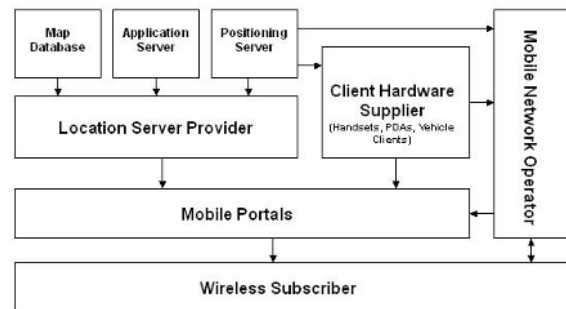
Managers can also choose to receive alerts when a worker arrives at a given location, or enters or leaves a specific location.

LBS VALUE CHAIN

The location-based services value chain consists of a well-defined ecology of infrastructure and service providers (see Figure 1). The members of this ecology are the wireless subscriber, mobile network operator, client hardware provider, mobile portal provider, and the location server provider cluster, which includes the map database, application server, and positioning server functions. This value chain is intended to provide subscribers with position-related services that combine wireless communication capability with network-based services and

Location-Based Services

Figure 1. The location-based services value chain (Adapted from Jagoe, 2003)



position information. The goal is to add timely, precise, useful location information that surrounds the user's position. For example, ESRI, a large GIS company, offers location-enabled services in a value chain that consists of 16 application developers and four major network operators.

The LBS value chain is continually evolving. Providers need to support a complex environment. Operator platforms comprise multiple technologies and multiple client devices that must be backward-compatible in many instances. LBS providers face high integration, development, and support costs to deal with this environment.

The Mobile Customer

The foundation member of the location-based services ecology is the mobile customer. No service will be adopted if the customer's value requirements are not met. Location-based services have suffered from a slow take up due to the lack of perceived value, the limitations of the technology, and the premium pricing imposed by the operators for data services.

Encouraged by government mandates for E911, safety has been targeted as a primary value driver in the form of emergency services that would prime the market (Dobson, 2003).

The initially poor renewal rate for GM's OnStar service was often cited as a major challenge to this assumption. In 2003, GM changed its advertising to "people freaking out" and the tide changed. Real-life situations where people were in accidents, locked out of their cars, or injured led to new customers and increased renewals (Masnick, 2004). This behavior may not carry over to handsets in the general case since people do not get lost that often and voice communication already performs essentially the same safety role.

The poor quality of location data is another barrier to growth. Location data that is not tied to a geographically relevant map display is clumsy to use and difficult for consumers to master. Some forecasters believe that "push-based" services will solve this problem since position-oriented information will automatically be cued by the user's location. Such services raise questions of interoperability and platform standardization that have yet to be resolved.

Perhaps the biggest barrier to consumer adoption of location-based services is the privacy issue (Dobson, 2003). Consumers have to be convinced that their personal location data will be kept confidential and only available for consumer-approved uses. For instance, will location-based personal information be made available to marketers, taxing authorities, plain-

tiffs' lawyers, or prying spouses? Several well-publicized compromises of personal data have convinced many customers that service providers have a poor record in this area. This perception must be overcome if customers are to embrace location-based services.

The jury is still out on how rapidly customers will adopt location-based services. If the history of mobile phone adoption is any indicator, then the benefits will have to be much more apparent and the prices much lower before the perceived value of this type of service is attractive for the mass market (Banforth, 2004).

The Mobile Network Operator

From the customer's perspective, the mobile operator sits at the top of the location-based services food chain. Customers form their relationships with the operator in most cases. The network operator is the gatekeeper that determines which members of the value chain will be invited to provide services, often under private-label conditions. The reason for this is the operator owns the subscriber interface and the customer relationship. All other players in the value chain are in a derived-demand position. Some exceptions exist, such as Yahoo, which already has substantial market power that is derived from its large subscription base.

The primary desire of the operator is to recoup its investment in the new 3G infrastructures, reduce churn, and to drive ARPU through the selling of new data services. As voice services have commoditized, ARPU has fallen. The common assumption among cellular operators is that the high use of these data-based applications will not occur until 2007-2008 (Miller, 2005). Even then the ARPU enhancement may be modest since most potential customers will wait for the prices to fall.

Mobile Client Hardware Providers

The mobile client hardware providers are key players in the value chain. They provide the hardware devices and other infrastructure that subscribers use to access location-based services. These devices include smart phones, PDAs, mobile computers, and vehicle systems (Jagoe, 2003). Companies such as Nokia, Motorola, Sony Ericsson, Palm, HP, LG, Samsung, and Qualcomm are examples of this type of provider. Operators depend on these companies for the R&D for handsets and infrastructure for the network. For location-based services the providers must not only tie to the network protocols, but also must work with other providers to ensure their hardware enables interoperability and application integration. The hardware providers are subject to the requirements, specifications, and private-label branding strategies of the operator.

Mobile Portals

Mobile portals are on-demand browser-based information management applications that extend location-based applications to mobile clients. Application aggregation and integration is the key to a successful portal service. The mobile portal service providers are at the forefront of this trend (Senia, 2005). Mobile portals can be entertainment based, such as those that focus on providing ring tones, music, video, or games. They can be universal portals that provide news, financial data, weather, and mobile commerce. Location portals typically can provide navigation, location-based information, location-sensitive transactions, and emergency services.

Mobile portals come in four general categories: operator-owned, semi-independent, independent, and Web incumbent (Grossman, 2001). Operator-owned portals, such as NTT

Location-Based Services

DoCoMo's i-Mode and Orange, are the first sites that users see. Operator revenues support them. Semi-independent portals such as T-Motion (T-Mobile/T-Online/Deutsche Telecom) and Vizzavi (Vodafone/Vivendi Universal) are joint ventures and may enjoy preferential status with the operators due to the existing relationship.

Pure-play independent portals such as Speedy Tomato and Zed offer ring tones, travel, weather, financial, and other services. Still others such as AvantGo (Sybase) and Breathe have ported Internet services to mobile devices that include phones and standalone PDAs. Internet incumbents include Yahoo, Google, AOL, and MSN that are established brands that have their own customer base and market WAP versions of their Web sites. They still have to contend with the operators' walled gardens. If the operators are able to continue their walled garden strategy, it is likely that the semi-independents and pure-play portal models will fail. Without the backing of the operators or existing Internet business models, they are exposed to considerable financial risks since operators rarely share their traffic revenues.

Although mobile portals serve as the important aggregation level of the value chain, other players can affect their success. Relatively few device makers dominate the market and can dictate technology. Operators want their own portals to have preference.

Location Server Provider Cluster

The location server (Autodesk, Webraska) provides location logic to integrate location service applications for mapping, visualization, location analysis, and positioning capabilities. The map database (Navtech, TeleAtlas, MapQuest) provides maps, routes, points of interests, and the locations of facilities such as hospitals, restaurants, fillings stations, and retail stores. Posi-

tioning services (Nokia, Intrado, Cellpoint) called gateway mobile location centers (GMLCs), provide fast and accurate location fixes subject to handset capabilities and the sophistication of the operator's network.

The application server (IBM, Oracle, BEA, Microsoft) level of the value chain provides the server-side platform for developing and deploying the location business application logic. It consists of a business logic layer, data access layer, and presentation layer (Jagoe, 2003). It is the core IT element of the value chain. It must integrate with the network infrastructure and client devices.

The location server cluster players provide "white-label" location solutions for wireless operators. Although the application-side players are still relatively small in terms of size, their contribution to value is essential to the ability to deploy and support location services.

Value Chain Dynamics

The industry structure and growth would depend on the dynamics and uncertainties in this sector, nature of technology, and the level of cooperation between the network operator, marketer, and other players in the mobile value chain. Industry consolidation at the operator level is increasing competition among value chain members as fewer, but larger, operators need fewer LBS function providers. In addition, larger providers, such as Autodesk, are integrating broad functionality into a suite of location applications that squeeze out smaller, less-integrated suppliers.

The "walled garden" strategy of the operators provides few opportunities for LBS providers to go directly to the subscriber. In most cases, providers are lower in the value chain than operators and must depend on them for business.

In this chapter, we restrict our focus to the discussion of issues surrounding the network operator (carrier) and the marketer/provider of location-based services. The roles of application developers, aggregators, and device makers are important, but beyond the scope of this chapter. Three different scenarios are discussed: network operator-dependent, network operator-assisted, and network operator-independent.

Network Operator-Dependent Scenario

The network operator-dependent business scenario prevails in the voice-based mobile services sector where the network operator is the dominant entity. Operators are heavily dependent on voice revenue. Operators view data services, and especially m-commerce data services such as LBSs as significant strategic opportunities for increasing revenues and profits.

With location-enablement in place, the network operator would generate and own the location data. The network operator would leverage this to provide location-based services or make this information available to marketers and providers of such services. Location-based service providers would enter into non-exclusive relationships with one or more competing network operators to obtain location-based customer data. Customers would only have access to services that are offered by their network operator. This scenario is similar to the existing WAP-based “walled garden” market structure.

Network operators can avail themselves of several recurring revenue streams under this scenario including LBS subscription fees (e.g., a monthly fee for roadside assistance service), per-usage fees for information services (e.g., directions to a restaurant), air time usage, and

packet-usage for the service transaction. Revenue would also result from revenue sharing, typically a transaction fee, with marketers of location-based services. In some instances, operators could also charge marketers and advertisers a fixed fee for content placement. This is similar to existing models where content providers pay network operators a fixed fee to get access to the network’s customers. Some location-based services may also be provided free or as part of a bundled service. This would help reduce churn, create differentiation, and increase market penetration.

Data mining of location-based customer data could potentially provide an additional revenue stream. Location-based services will generate extremely large data sets from customer transactions. Location coordinates, purchase records, financial data, travel patterns, customer service records, and m-commerce sites accessed are just some of the data that will be collected and mined (Harmon, 2003). Operators that have superior customer knowledge would also have a distinct advantage in attracting providers of location-based services and other services. This should result in better-designed services that would help network operators build stronger relationships with customers, reduce churn, and increase revenues.

In this market structure, the operator provides end-to-end value. In some cases when the brand identity of the location-based service is not well established, the operator could re-brand the service as its own and create a point of differentiation in the market. Operators gain competitive advantage by associating or co-branding services with well-established brands. Location-based services provide operators with a way to enhance the value proposition to their customers. It could provide a means to address a major criticism of the “walled garden” approach arising from lack of services and limited content (Sharman, 2001).

Location-Based Services

For marketers of location-based services, network operators provide the marketing channel that gives them access to a captive customer base. They derive revenue by charging subscription or per-usage fees to customers, and/or by selling content to the network operators. To increase their customer base, marketers would have to partner with multiple operators.

For customers, the major source of value would be availability of potentially useful, and potentially life-saving, services through a known entity, the network operator. Availing themselves of premium location-based services would typically be more convenient and secure, and engender more trust when provided by the network operator, rather than purchased directly from an unknown LBS company. Customers also have the convenience of dealing with a single bill from the operator. However, customers in this scenario would be unable or might find it expensive to go outside the “walled garden” to access services not offered through the network operator.

Network Operator-Assisted Scenario

The network operator-assisted business scenario is characterized by the network operators’ relinquishment of exclusive control over location data. Location data would become available to a location-based service provider either free or at a given rate. Palm.net provides approximate location data at no charge by making available the zip code of the wireless tower nearest to a Palm PDA user to any application (location service) that requests this information. Users can access information about retail establishments in a given area.

Network operators tend to become the “dumb pipe” or mere transporters of voice, data, and other content under these conditions. They generate revenues through selling loca-

tion data and through increased airtime usage and/or data packets volume. Network operators may give up control over location data when they are unable to provide meaningful content to their customers. By opening the network and relinquishing control over location data, operators would attract more marketers of these services to the network. Marketers would have the potential to capture greater value, because they would be dealing directly with customers. Customers would benefit through greater choice and the ability to deal directly with marketers. Market forces would determine which services survive in the marketplace. The lack of a consistent billing aggregation infrastructure could be a source of confusion and an inconvenience for the customers. However, significant progress has been made with prepaid and credit card billing platforms, e-wallets, and more robust carrier billing platforms brought about by the recent industry consolidation (Kalakota & Robinson, 2002).

Network Operator-Independent Scenario

Location data does not always reside or get generated on the cellular network. Pure hand-set-based solutions such as those that rely on GPS technology for generating location data are network independent. Stand-alone networks have targeted market niches such as those being addressed by companies such as Cell-Loc for tracking truck fleets, assets, traffic, children, and pets, or by LoJack for recovering stolen cars. In addition, mobile devices that are enabled for use in “hot spots” of wireless LAN networks, as well as Bluetooth-enabled devices, would also fit in this scenario.

In some cases the LBS marketers simply rent capacity on the cellular network. In this case the LBS marketers of the location-based services capture most of the value since the

customer does not know or care which network is being used for the service. The network operators realize revenue by providing transport for the location-based services. Independent services could be subscription based, charged on a per-usage basis, or both. Customers exert choice preferences by seeking out a different service provider. Billing is direct to the service provider and not generally aggregated with other services.

Currently, wireless LAN operators are providing fast Internet access to customers. In the near future, the potential for providing location-based services, especially promotions and advertisements, could be realized when these networks are set up in malls and other areas with a high density of retail stores. The marketers (e.g., Sears) would benefit by offering personalized advertising and promotion to customers and increasing traffic to their stores. Customers would have to use mobile devices with appropriate hardware and software. Similarly, Bluetooth-enabled devices would also be an appropriate vehicle for location-based advertising. Network operators could capture some value by aligning with wireless LAN operators and marketers to increase coverage and access to customers.

STRATEGIC ISSUES

The emerging brand strength of the content and application service providers will challenge the power of the network operators. However, even in this changing environment, operators have considerable power because they own and therefore can leverage the location data generated in their networks, its authentication information, user preferences, and the billing relationship with their customers (Prem, 2002).

Developing Content and Relevant Services

Operators would capture most value by keeping and using the location data to develop tailored services in-house. The services they provide would be a source for differentiation in the market. However, developing good content for services is not their forte. Alliances with specialist content providers would therefore become necessary. The nature of these alliances would be an important strategic issue for operators.

A “walled garden” approach, similar to strategies used currently for wireless voice services, is one option. This approach offers operators a high degree of control and allows them to retain a greater proportion of the revenues. The operators could either brand the content/service as their own house brand or co-brand the service depending on the brand image of the service provider. For example, a network operator could brand a traffic update service as its own by selling or sharing location data with a service provider that provides location-specific traffic updates. On the other hand, if the service is being provided by a well-established brand like the American Automobile Association (AAA), then a co-branded approach may work to the network operator’s advantage. It also allays privacy concerns of customers, presumably because the network operators would share location information only with trusted partners (Akçayli et al., 2001). However, the range of content and services would likely be limited, and content developers may lack the incentive to market their content or be a part of the walled garden (Sharman, 2001).

Alternatively, an “open garden” approach implies a scenario where content developers and marketers have access to location data and deal directly with consumers. The network

Location-Based Services

operators could develop a shared revenue model with their partners. The content developers and marketers would be free to set prices for their services and attract customers. The major concern for network operators would be reduction of their role to mere channel providers for various services. However, the ownership of the billing relationship with the LBS customers would provide operators with some leverage to negotiate favorable revenue-sharing agreements (Sharman, 2001).

Service providers and LBS marketers have numerous choices concerning the type of technology to use for location-based services. This would have significant implications on how they realize revenues. They would be dependent on the network operator for location information if they used the mobile networks to reach customers. However, obtaining location information from networks provides these service providers with the ability to reach a large base of consumers. Their ability to realize revenues from services would be dependent on the agreement they have with network operators. If location data were generated outside the network (such as through GPS systems), they would potentially have more flexibility in the creation and delivery of services. It is imperative for service providers and marketers to identify potential technology winners in this uncertain domain. For marketers with physical stores, WLANs or Bluetooth technology appear to be more appealing alternatives. However, it may take some time until the technology-enabled consumer base is large enough for the LBS business model to be effective.

Pricing

Arriving at a successful pricing strategy for LBS is another key issue for carriers that can either inhibit or facilitate market growth. Consumers should be able to see the link between

the value they receive from location-based services and the prices they are charged (Sharman, 2001). The network operators were badly burned by the failure of WAP-based information services to generate profits. It is therefore critical to link marketing strategy to the creation of customer value with these services.

Perceived value is conceptualized as the overall benefit derived from a service, at the price the customer is willing to pay (Harmon & Laird, 1997). The benefits of location-based services include personalized, context-specific, timely, convenient services. The costs include fees and other charges for the service, as well as non-monetary costs such as loss of privacy and intrusive marketing promotions.

Mere availability of a wide number of services may not be perceived as valuable. Segmenting the market and differentiating location services by perceived customer value should lead to improved financial performance. For example, roadside emergency services such as OnStar, which can guide a driver in distress, may be very important for frequent travelers and soccer moms, but not for college students. Any attempt of a "one-size-fits-all" approach would likely fail (Williams, 2003). Applications developed for niche segments such as outdoor enthusiasts like campers, hunters, and fishermen may hold considerable potential (Rao & Minakakis, 2003).

They have typically use-per-minute and flat-rate charging models for voice services, but few customers have been willing to sign up for the open-ended pricing structure for data services. The new higher-speed 3G networks will present new opportunities to, hopefully, get the price model right. At present, consumers will most likely be charged on the basis of airtime and the quantity of data transferred, which has not worked well for the mass market. Other payment options may include a "pay-per-use," "all-you-can-eat," prepaid usage; service-spe-

cific subscription fees; time-of-day rates; and volume rates (Prem, 2002).

Network operators have to evaluate the trade-off between benefit to the customer and cost to the carrier with each pricing option. In the U.S. market, consumers are used to flat-fee tiered pricing (where a certain quota of airtime is included for a given fixed monthly charge). A switch to pricing by quantity of data transferred is not an easy pricing model to sell to consumers since it has the ability to generate unexpectedly high bills. Subscription-based pricing may, on the other hand, be easy to communicate but may have limitations on how much value the carriers could capture, especially if the customer is a heavy user of the service.

For marketers and service providers, there would be most flexibility in pricing matters when consumers have independent access to their services without restrictions or charges from the network operators. They would have the least flexibility when operating in a “walled garden” type of scenario or in situations where their services are re-branded as the house brand of the operator. In an “open garden” approach and other variations of this approach, the entity that pays for the ad transport charges would have an impact on pricing for services. Even if the ad-transport charges are picked up by the marketer, the issue of pricing is relevant. Marketers who plan to use location-based advertising and promotions to drive traffic to a storefront or encourage mobile-shopping behavior are also faced with serious pricing issues. Will customers be willing to pay for the LBS? Strategies to attract customers to their stores are untested, especially in terms of personalized promotional offers and pricing of bundled services.

Marketers could potentially adopt some form of first-degree price discrimination and charge a highly customized price that corresponds to their willingness to pay for that service based

on the perceived value of that service. Further, a justification could be provided for charging different prices based on location, time, and context. This approach may be possible in m-commerce situations because the customer is at the location, he or she is ready to buy, and price comparison information may not be freely available to consumers in this context. In contrast, in the e-commerce and traditional retail formats, it is difficult to implement this type of value-based pricing because price comparison information is more readily available (Subramanian, Peterson, & Jarvenpaa, 2002). However, in situations where marketers and service providers are dependent on the network operators, they may benefit from aligning themselves with network operators and offering their services as part of a bundle of services from the operator. This would ensure a higher penetration for these services in the short term while raising brand awareness.

Addressing Privacy Concerns

Acceptance of location-based services would depend to a great degree on how well operators (and service providers) allay privacy concerns of consumers. It is critical for operators to build trust with their customers. It is widely agreed that unsolicited mobile advertising is viewed as intrusive and unwelcome. Other services such as being located on a tracking service may raise security and privacy issues. Therefore, operators should carefully evaluate the option of making these services opt-in and permission based (Newell & Lemon, 2001). Initial implementations of LBSs such as Bell Mobility’s MyFinder, TeliaSonera’s Friend Finder, and AT&T’s Find Friends demonstrate that operators are cognizant of privacy issues and have taken appropriate steps to have sophisticated authentication and authorization frameworks (Spinney, 2004). Furthermore, developments in

Location-Based Services

the regulatory side such as the CAN-SPAM Act of 2004 has provisions that prohibit the transmission of unsolicited wireless messages (Carmody, 2004). However, privacy regulations are considered far more stringent in Europe, where the principle of explicit opt-in is adopted (Camponovo & Cerutti, 2004).

Marketers need to recognize the fine balance between personalization and addressing privacy concerns. Truly personalized context-specific LBSs would require customer data from several sources. Marketers could potentially access a customer's financial and off-line shopping data, which is then used for tailoring messages and services for that consumer. When the use of personal offline data along with location information becomes apparent to the consumer, serious privacy concerns may arise (Subramanian et al., 2002). Established retailers and marketers may have an advantage over unknown or fledgling providers of services. However, they could quickly erode this advantage if they mishandle privacy concerns of their customers.

While the E911 mandate requires an infrastructure that provides fairly precise location data, operators should take into consideration the level of precision desired to offer a service. Some experts envision a scenario where network operators would have a tiered relationship with content providers and marketers. They would share precise location data with only trusted partners (Akçayli et al., 2001). Marketers may also be better off if they trade-off gains from a high degree of personalization to securing their consumers' privacy (Pfenning, 2001). Services related to emergencies would require precise location data, but services that provide a list of restaurants in a neighborhood may require less precise data. Marketers would have relatively greater ease in having access to less precise location data than very precise location data.

There is evidence from e-commerce that users are willing to provide personal information in exchange for personalized services (Beinat, 2001). A recent study also reveals that mobile customers are more concerned about usefulness and degree of personalization of service than privacy threats (Ho & Kwok, 2003). However, the potential for misuse of personal information is real and raises several questions such as: Who would have access to the location information? Can the user not have his location tracked? How secure is the information (Fink, 2002)? It is important for the industry to be proactive in instituting good practices regarding the management and use of location data. This not only builds trust with consumers, but it also serves to avoid a more restrictive regulatory climate.

Lack of Interoperability

Lack of interoperability between the wireless network platforms and with the location-awareness technology will continue to be a major issue well into the 3G-network rollout. This has a direct impact on market growth because it affects software, device, and content development. It also has an adverse impact on customers due to high costs, inconsistent user interfaces, uncommon modalities, and steep learning curves access and use the new services.

Another potential barrier to growth for LBSs and m-commerce in general is the lack of standardized mobile payment infrastructure. As new modes of payment (e.g., charging for content, charging by usage, pre-paid usage) emerge in this market, it is vital that appropriate payment structures are in place to facilitate easy-to-use, secure, flexible payment methods. Transformation of the mobile device into an electronic wallet would be an important facilitator of overall growth in m-commerce. The mobile phone/device as a means of payment

Table 3. Strategic issues for the commercialization of location-based services

VALUE-CHAIN PARTNER	STRATEGIC ISSUE
Network Operator	Develop high-value LB services Walled garden vs. open garden Structure of alliances and partnership with LBS providers Choice of LBS technology Increase user base and ARPU Choose pricing/revenue models Data ownership Privacy/security concerns
LBS Provider/Marketer	Develop high-value LBS applications and content—map value drivers Choose revenue and pricing models/fee structures Create alliances and partnerships with carriers/network operators and other content and application providers Online/offline integration Privacy/security concerns

would be critical in facilitating impulse decision making sparked by location-based services (Kalakota & Robinson, 2002).

The major issues for operators and service providers are summarized in Table 3.

CONCLUSION

The continuing rollout of the technological infrastructure that enables the generation of location-awareness data is regarded as a major growth opportunity for wireless data services. The mobile device that is used primarily for voice services now is expected to morph into a comprehensive, multi-purpose personal device. The translation of market opportunity for location-based services to increased revenues and profitability depends on how the different players in the value chain resolve the key strategic issues they are facing.

Creating high-value services that can generate strong interest with potential customers is the top priority for this industry. Network operators dominate the value chain for location services. In Japan and South Korea, this domi-

nance has led to relatively quick deployment of location-based services that have been successfully adopted by consumers and businesses. In Europe and North America this operator-centric model has not yet led to success. The subscribers' location-based services needs are poorly understood. The value-creation aspects of the business model remain in doubt.

Providers need to support a complex environment. The lack of standardized location platforms increases development and support expenses. Operators want to offer proprietary offerings. The end game for mobile location service providers is not encouraging if the operators continue to have their way. Future profits for map data, location data, positioning, and comprehensive location service providers will be constrained by the operators. The consumer uptake of these services will continue to be slow, as prices remain high. The solution to this situation is for the operators to create a more standardized hardware and software environment, open up their platforms, and otherwise enable those companies that can create compelling location-based services.

REFERENCES

- Akcayli, E., Brooks, D., Laszlo, J., & McAteer, S. (2001, November 20). Location-aware applications—Improving end-user value despite carrier hesitation. *Jupiter Research*. Retrieved June 3, 2002, from <http://www.jupiterdirect.com/bin/report.pl/89027/1057/>
- Angelides, J. (2005, January 26). U.S. poised to capitalize on location services. *Directions Magazine*. Retrieved August 20, 2005, from http://www.directionsmag.com/article.php?article_id=739&trv=1
- Banforth, R. (2004, June 17). *Mobile location services: Dead reckoning or dead duck?* Retrieved October 20, 2004, from <http://it-analysis.com>
- Beinat, E. (2001, September). Privacy and location-based service. *Geo Informatics*, 14-17.
- Blackwell, G. (2001, November 16). *Location-based services in the WLAN world*. Retrieved November 10, 2003, from <http://www.80211-planet.com/columns/article.php/925511>
- Burnham, J. P. (2002). *The essential guide to the business of U.S. mobile wireless communications*. Upper Saddle River, NJ: Prentice-Hall.
- Camponovo, G., & Cerutti, D. (2004, July 12-13). The spam issue in mobile business: A comparative regulatory review. *Proceedings of the International Conference on Mobile Business*, New York. Retrieved from <http://www.hec.unil.ch/gcampono>
- Carmody, B. (2004). *Wireless promotions: Future vs. fantasy*. Retrieved December 11, 2004, from <http://www.seismicom.com/WirelessPromotions.pdf>
- Christensen, C. M., & Tedlow, R. S. (2000, January-February). Patterns of disruption in retailing. *Harvard Business Review*, 78(1), 6-9.
- Dobson, M. (2003, March 26). Where will the money flow in the United States' location based services value chain? *Directions Magazine*. Retrieved February 5, 2004, from http://web2.directionsmag.com/article.php?article_id=311&trv=1
- Faggion, N., & Trocharis, A. (2004). Location-based services strengthen the strategic position of mobile operators. *Alcatel Telecommunications Review*, (4th Quarter), 2-8.
- Fink, S. (2002). The fine line between location-based services and privacy. Public safety report. *Radio Resource Magazine*, 17(7). Retrieved February 5, 2004, from http://www.jlocationsservices.com/LBSArticles/Sun_RadioResource.pdf
- Grossman, W.M. (2001). Diminishing mobile portals. *The Feature: It's All about the Mobile Internet*, (October 10). Retrieved from www.thefeature.com
- Harmon, R.R. (2003). Marketing information systems. In H. Bidgoli (Ed.), *Encyclopedia of information systems* (pp. 137-151). St. Louis, MO: Elsevier Science.
- Harmon, R. R., & Laird, G. (1997). Linking marketing strategy to customer value: Implications for technology marketers. In D. F. Kocaoglu, T. R. Anderson, K. Niwa, D. Milosevic, & M. J. Gregory (Eds.), *Innovation in technology management: The key to global leadership, PICMET Proceedings* (pp. 896-900). Piscataway, NJ: IEEE.
- Ho, S. Y., & Kwok, S. H. (2003). The attraction of personalized service for users in mobile commerce: An empirical study. *ACMSIGecom Exchanges*, 3(4), 10-18.

- Jagoe, A. (2003). *Mobile location services: The definitive guide*. Upper Saddle River, NJ: Prentice-Hall.
- Kalakota, R., & Robinson, M. (2002). *M-business: The race to mobility*. New York: McGraw-Hill.
- Masnick, M. (2004, June 16). Telematics creeping out of the dashboard. *The Feature: It's All about the Mobile Internet*. Retrieved from www.thefeature.com
- Miller, L. (2005, February). Enhanced services: As IP technology takes over, enhanced services will finally attain the popularity they deserve. *Von Magazine*, 32-33. Retrieved March 25, 2005, from <http://www.vonmag.com/>
- Newell, F., & Lemon, K. N. (2001). *Wireless rules: New marketing strategies for customer relationship management anytime, anywhere*. New York: McGraw-Hill.
- Paavilainen, J. (2002). *Mobile business strategies*. London: IT Press.
- Pfenning, A. (2001, August 27). Personalization: Delicate balance. *InternetWeek*, 40-42. Retrieved March 20, 2003, from <http://internetweek.cmp.com/eresearch01/data082701.htm>
- Prem, E. (2002, February). Innovative mobile services and revenue models. *EUTEMA Report*. Retrieved March 20, 2003, from <http://www.eutema.com/eutema-Report2-02.pdf>
- Rao, B., & Minakakis, L. (2003). Evolution of mobile location-based services. *Communications of the ACM*, 46(12), 61-65.
- Reid, H. (2004, August 27). Location-based services—we're getting closer: Lessons learned from ESRI's LBS summit. *Directions Magazine*. Retrieved December 10, 2005, from http://web2directionsmag.com/article.php?article_id=648
- Robinson, T. (2000, September 13). Wireless marketing is about location, location, location. *InternetWeek*. Retrieved March 20, 2003, from <http://www.techweb.com/wire/story/TWB20000913S0014>
- Senia, A. (2005, March 1). *Broadband and wireless spark U.S. telecom growth*. Retrieved from <http://www.americasnetwork.com/americasnetwork/article/articleDetail.jsp?id=147750>
- Sharman, D. (2001, March). *Pricing models for wireless Internet services*. Retrieved from <http://www.wmode.com>
- Spinney, J. (2004, March 1). Location-based services and the proverbial privacy issue. *Directions Magazine*. Retrieved December 10, 2005, from http://news.directionsmag.com/article.php?article_id=510&trv=1
- Subramanian, S., Peterson, R. A., & Jarvenpaa, S. L. (2002). Exploring the implications of m-commerce for markets and marketing. *Journal of the Academy of Marketing Science*, 30(4), 348-361.
- Williams, D. H. (2003). *It's the (LBS) applications, stupid!* Retrieved from www.e911-lbs.com

Chapter III

Location Services in Cellular Networks

Israel Martin-Escalona

Technical University of Catalonia, Spain

Francisco Barcelo

Technical University of Catalonia, Spain

ABSTRACT

This chapter presents the main features entailed in providing location services in cellular networks. It begins by describing some of the most important location-based services and the main location techniques that have been developed in order to allow these services to be provided. These techniques involve several constraints that reduce their applicability in certain environments. Several of these restrictions are explained in this chapter, as well as the solutions proposed in order to overcome them. Regulator bodies have included some of these techniques in their official recommendations. This chapter also reviews the location architectures standardized for use in the main cellular networks and presents the concept of location middleware as a natural addition to these architectures.

INTRODUCTION

Location Services That Make Use of Customer Position

Mobility is a key factor in the provision of services in cellular networks. Although most of

the technical limitations (error rate, bit rate, etc.) of 2G and 2.5G networks were overcome in 3G networks, services seem not to have taken off. Most experts agree that in order to revitalize the services market, new wireless services that represent a real addition of value must be developed. And this does not mean

merely translating services from the wired to the wireless context, but developing new value-added services for this particular environment. One example of this set of new services is the location-based service (LBS). LBSs are services that adapt to a user's location and situation—that is, applications that are dependent on a certain location. Goodchild (2001) states: "LBSs exploit the ability of technology to know where the user is and to modify the presented information accordingly." All definitions agree that LBSs are services that in some way make use of the knowledge of a user's position. Currently, many LBSs have already been designed and are ready to be implemented for commercial exploitation. A selection of the more relevant ones is described as follows.

- **Information Services:** These provide the location of the nearest entity that the user is interested in. Requesting the address of the nearest gas station or the cinema that is closer to our location would be examples of these kinds of services.
- **Navigation:** These services guide the user to a certain destination. A typical example of this service is the navigation system included in the highline cars which leads us to a certain destination.
- **Workforce Management:** This includes any service that can be used to manage a workforce, such as assigning tasks in real time according to the position of the workers and the location of the task. For instance, a company focused on support and maintenance may assign the issues according to the position of their workers, thus reducing the delivery costs.
- **Demand-Responsive Transport:** These services are transport services that have been upgraded with position information, which allow transport companies to provide their services according to customer

location or destination. One example of this kind of service is the management of a taxi fleet, where the position of the customer helps to select the taxi that has to take the order.

- **Lone Worker Applications:** These services aim to assist workers that may be involved in dangerous situations, such as firemen and policemen.
- **Children Tracking:** These services provide the position of children in real time or alert the customer when a person being tracked ventures beyond certain boundaries.
- **Medical Alert:** These kinds of services are aimed at people who are frequently ill. For example, the elderly can alert the nearest medical center in the case of illness and thus provide information about their situation.

Features and Classification of Location-Based Services

LBSs may be classified in several ways. From a business point of view, location services can be sorted according to three criteria (see Table 1):

1. **Service Purpose:** According to its value for the customer, such as what kind of utility it offers. This group includes several types of LBSs, such as tracking and monitoring, assistance, and location-based information.
2. **Customer Participation:** As customers may adopt an active or a passive role during the execution of the location service, there are accordingly two different types of LBSs: push and pull services. The user actively runs push services. Pull services, however, are not initiated by an explicit customer command; instead, they

Table 1. Location-based services classification (Deirmentzoglou, 2004)

Trigger Services E-commerce, payment information, advertising, etc.	Tracking and Monitoring Fleet management, telematics, asset tracking, etc.	Location-Based Information Traffic and navigation, entertainment, mapping, etc.	Assistance Services Personal/vehicle emergency, roadside assistance, alarm management, etc.
Push Services Travel directions, taxi hailing, m-commerce, etc.		Pull Services Zone alerts, traffic alerts, etc.	
B2B Fleet and freight tracking, etc.	B2C Discounts, ads, special events, etc.	C2C Find a friend, primary schools, etc.	C2B Find a gas station, community events, etc.

are triggered according to predefined rules that depend on the position of the user.

3. **Business Model:** According to the market to which it is addressed—that is, business to business, business to consumer, consumer to consumer, and consumer to business.

Location services can also be classified according to technical parameters. In the case of LBSs, quality of service (QoS) parameters are usually proposed. Several parameters may be used to measure QoS in LBSs. In 3GPP TS 03.71 (2002) and 3GPP TS 23.271 (2004), the 3GPP proposes quantifying QoS according to

two parameters: minimum accuracy and maximum response time. Minimum accuracy refers to the maximum error that an LBS supports. Usually, customer position is delivered to the location client as a pair: position and possible location area. The position represents the estimated position, while the location area boundaries set the limits of possible error range. Maximum response time is defined as the maximum time that the LBS takes until its completion. These two parameters are included in all the QoS definitions for LBSs. However, several parameters may be added, depending on the scenario. For instance, availability indicates the coverage expected by the user (i.e., the

Table 2. Quality of service required by some LBSs in different scenarios (Wilde, 2002)

LBS	Indoor		Urban		Suburban		Rural	
	Accur	Yield	Accur	Yield	Accur	Yield	Accur	Yield
<i>Information Services</i>								
<i>Basic</i>	50-100m	80.0%	50m	80.0%	1000m	80.0%	5000m	80.0%
<i>Enhanced</i>	20-50m	90.0%	50m	90.0%	50m	90.0%	50-100m	90.0%
<i>Navigation</i>	10m	95.0%	10m	99.9%	10m	99.9%	20m	99.9%
<i>Worker Management</i>	50m	95.0%	50m	95.0%	50m	95.0%	100m	95.0%
<i>Worker Tracking</i>	50m	99.9%	50m	99.9%	50m	99.9%	100m	99.0%
<i>On-Demand Transport</i>	n/a	n/a	50m	90.0%	50m	90.0%	50m	90.0%
<i>Children Tracking</i>	50m	99.9%	50m	99.9%	50-100m	99.5%	50-100m	99.9%
<i>Medical Alert</i>	50m	99.9%	50m	99.9%	50-100m	99.9%	50-100m	99.9%

percentage of time and space in which the LBS is operative). Table 2 shows the requirements for several LBSs. In this table, *Accur* and *Yield* stand for minimum accuracy (i.e., maximum error) and coverage, respectively.

Quality-of-Service Regulation

The QoS of certain LBSs should not depend solely on the service providers; rather, they should be regulated by standardization authorities due to their special features. Emergency services are the clearest example, since they are designed to provide the position of customers in critical situations. This constrains the minimum QoS that should be had when positioning mobile users. The United States was the first country to implement the infrastructure needed to manage emergency services. These services, known as 911 services, were regulated by the FCC (Federal Communications Commission) and were initially designed to operate in wired networks that allowed customers to be located with a sufficient degree of precision. However, the application of these services to wireless networks was subject to two significant limitations: wireless networks did not provide the ALI (automatic location identifier) or the ANI (automatic number identifier). The ALI is used to determine the geographic position of the user, and the ANI provides the information needed to call the user back if necessary. The FCC updated the regulations for 911 services and created the E-911 (Enhanced 911) service to overcome those limitations. In the initial stage (Phase I), all wireless-network operators were asked to provide call centers with ANIs and the approximate locations of users (e.g., identifiers of the cell). The second stage (Phase II) was aimed at providing accurate positions. At first, the FCC only allowed network-based location techniques to be used in providing ALIs. However, the

Table 3. FCC recommendations for accuracy in emergency services

Percentage	MS-based	Network-based
67%	50 m	100 m
95%	150 m	300 m

upgrading of location techniques since 2001 has meant that the FCC has had to update the E-911 rules again, including the use of mobile-based location techniques—that is, location techniques which calculate the position in the user terminal. Accordingly, Phase II set the QoS requirements for LBSs, as shown in Table 3. The E-911 rules also stated that, by January 2006, all the wireless network operators must be able to provide emergency LBSs that fulfill the figures set out in Table 3.

The EU, on the other hand, regulated the provision of emergency services through the E-112 regulations. The European Commission set up CGALIES for this specific purpose. This group had to define the requirements for the provision of location services in Europe, in such a way that this position information could be shared with the E-112 emergency service. The outcome was a document (Malenstein et al., 2002) in which the group put forward several proposals for achieving the integration of location information and emergency services in Europe. Accordingly, the EU upgraded the regulation of the provision of the universal service to include several notes on the provision of location information in E-112 emergency services. In fact, Article 26 of this recommendation states:

Member States shall ensure that undertakings which operate public telephone networks make caller location information available to authorities handling emergencies, to the extent technically feasible, for all calls to

the single European emergency call number '112'.

A text (European Commission Recommendation 2003/558/EC, 2003) that was published in the official *Journal of the European Commission* included several recommendations related to the provision of location information in the E-112 service, one of which was as follows:

Following on from the recommendation by CGALIES, providers of the public telephone network or service should use their best effort to determine and forward the most reliable caller location information available for all calls to the single European emergency call number 112.

Location requirements for emergency services in Europe are less restrictive than in the United States, and this situation is likely to continue until 2005, when the European Commission is to review the recommendations for including location information in the E-112 services.

LOCATION TECHNIQUES

Introduction

Quality of service in location services depends on several factors, such as the location technique used to retrieve the user's position, network infrastructure, performance of the user terminal, and so forth. However, the way in which the position is estimated seems to be the most important of these factors. For example, most of the time spent by an LBS is consumed in the position-achievement process. Accordingly, the selection of the location technique used to compute the customer's position will impact the performance of any location system and, consequently, that of the LBS being provided in it.

In Soliman and Wheatley (2002), geolocation is defined as "the process of determining the coordinates of an object on the surface of the earth." Many location techniques may be used to geolocate a mobile user. Location techniques can be classified according to several factors, one of which is the place where the position is really calculated, that is, where the *position calculation function* (PCF) is implemented:

- **Mobile-based techniques** compute the position in the user terminal. This may involve some information from the network (i.e., assistance data).
- **Mobile-assisted techniques** calculate the position in the network, even though the necessary measurements can be performed in the user terminal. Mobile-assisted techniques may also use information from the network.
- **Network-based techniques** compute the position in the network and use only data available in the network.

Location techniques can also be classified according to their QoS. The definition of quality of service for location techniques is a little broader than for location-based services. QoS can be defined according to the following four parameters (Soliman & Wheatley, 2002):

1. **Accuracy** is the maximum error provided by the location technique in the geolocation process. The accuracy figures for the location techniques will be compared to the requirements of the LBSs.
2. **TTF** stands for the response time (or latency) of the location technique—that is, the time required by the technique to geolocate the user. This parameter may become very restrictive for LBSs that need to periodically retrieve the user position. This could be the case of tracking services, used for instance by delivery

Table 4. Accuracy and TTFF for several location techniques

Technique	Network Impact	Terminal Impact	Accuracy	TTFF
Cell ID	None	None	250m-20km	< 1 sec.
Signal strength	None	None	100m-10km	< 1 sec.
TOA/TDOA	Medium	Low/medium	40-150m	< 1 sec.
AoA/DoA	High	None	50-150m	< 1 sec.
Fingerprint	High	None	50-150m	Seconds
GPS	Low	Very high	3-50m	Seconds
Hybrid systems	Depends on the techniques hybridized	Depends on the techniques hybridized	3-100m	Seconds
Ultra-wide band	Dedicated infrastructure	Very high	10-50cm	< 1 sec.

companies in order to locate the goods they transport.

3. **Yield** is the capability of the location technique to provide results even in difficult locations. This parameter should take a value of between 75% and 99% (Soliman, & Wheatley, 2002)—that is, the location technique should be available to at least 75% of the requests. Note that this percentage might be more difficult to achieve depending on the environment (e.g., in indoor environments).
4. **Consistency** is the capability of the location technique to produce location fixes independently from the environment and network technology. For example, would it be reasonable to think that a GPS (global positioning system) location technique would provide the same quality in a GSM network and a UMTS network? These kinds of questions are the ones addressed by consistency.

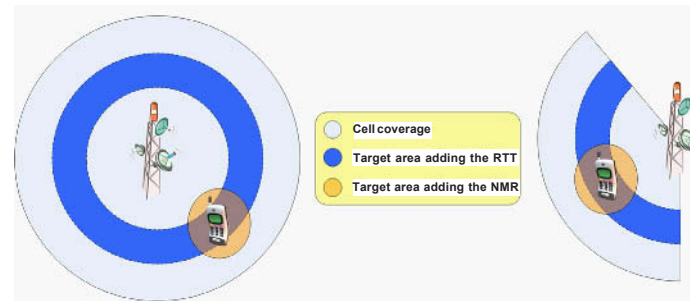
Some figures on the accuracy and TTFF provided by several location techniques are presented in Porcino (2001) and displayed in Table 4. More information on the location techniques presented in Table 4 is provided in subsequent sections.

Features and Performance

Cell Identification (Cell ID)

This technique geolocates the user according to the cell the user is located in: the position of a representative spot inside the cell is reported as the user location (e.g., the base station, the midpoint of the cell, etc.). The accuracy of such location is poor and obviously depends on cell size: since cells in urban areas are typically smaller than in rural areas, accuracy is better. The consistency of this location technique is also poor since the user has no knowledge of the cell size, and the accuracy may only fulfill the QoS of an LBS in specific areas. On the other hand, this technique presents good TTFF values since the knowledge of the cell is usually included in the location request itself. Furthermore, this technique presents excellent yield values since the only requirement is to be connected to the network. Today, many mass market services employ this technique, since it allows location-based services to be carried out in current legacy terminals and requires only minimum changes to the network infrastructure: cell ID is suitable for entry-level services.

Figure 1. From left to right, positioning with Cell ID++ in omni-directional and sectorized cells



Network Measurement Report (NMR)

This technique computes the distance between the base station and the terminal according to the attenuation introduced by the radio-path propagation. The data needed to perform the positioning are the power strength transmitted by the user terminal, the power strength received by the base station from the terminal, and a precise model to characterize the channel's radio losses (i.e., propagation, fading, attenuation due to weather, etc.).

The technique presents several limitations. For instance, although the transmitted and received power strength that are used are known in some networks (e.g., GSM), this information is not available in all systems. Another issue is that the user terminal is rarely within the base station's line of sight. Signal processing provides several mechanisms that limit the influence of this factor, but algorithms have an impact on the performance and the cost of the network equipment. Accordingly, the accuracy may be worse if the element that makes the power measurements in the network (usually the base station) is not able to completely remove the multipath effect. Furthermore, a very accurate radio model is necessary if the power is to be estimated with precision. This is

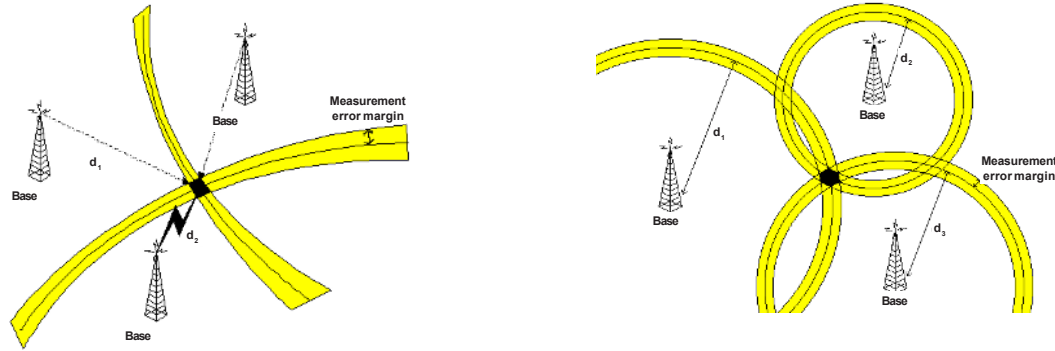
a hard task, since the model depends on the environment, which is highly variable (e.g., changes caused by factors such as rain, trees, buildings, or traffic jams are not known by the system).

These limitations lead to a poor consistency; thus, it is seldom used as stand-alone but rather combined with other techniques.

Enhanced Cell ID (Cell ID++)

This technique consists of upgrading the cell ID approach with additional information, such as the *RTT* (round-trip time) and the *NMR*. The *RTT* uses time advance information to reduce the location area. Thus, the target zone is reduced from the whole cell to the ring traced by the *RTT*. The *NMR* can also be used to limit the area in which the user is supposed to be placed. Figure 1 illustrates this procedure, in which different data are used to reduce the target area and thus achieve a more accurate position. Even with the addition of the *RTT* and *NMR* data, positioning using Cell ID++ shows poor consistency, since none of the combined techniques shows good consistency. However, the addition of extra information yields better accuracy.

Figure 2. From left to right, hyperbolic and circular modes of E-OTD (3GPP TS 03.71, 2002)



Uplink Time of Arrival (U-TOA)

The uplink TOA method is based on measuring a given signal sent from the mobile station and received by three or more measurement units. This method requires additional measurement unit hardware in the radio access network to accurately measure the TOA of the bursts. Since the geographic coordinates of the measurement units are known, the position of the mobile telephone can be calculated via hyperbolic triangulation. Usually, the burst that is used to locate the user is caused by a “virtual” handover request—that is, a request forced by the network to make the user terminal transmit the access burst. This method works with legacy terminals in most of the existing mobile networks with no need for modifications. However it has an impact on the performance of terminals, as the handover process is usually one of the most power-demanding processes.

Performance data presented in TruePosition (2004) show that this technology achieves accuracy figures in the range of 50 meters and fair consistency: this accuracy may be altered by several factors, including the signal-to-noise ratio (SNR) of received signals, the bandwidth of the transmitted signal, and the time available to the location processor to process the information from signals received by multiple antennas. The yield of this technique reaches 99%.

Enhanced Observed Time Difference (E-OTD)

E-OTD is a particular implementation of the OTD technique that consists of measuring the downlink TOA. This kind of location technique uses propagation-time values to compute the position of the user. The performance of E-OTD is as follows. All base stations transmit some burst periodically (there is no need for synchronization between base stations). Then, the user terminal calculates the difference between the times of arrival of the bursts from several base stations. In the event that the base stations are not synchronized, synchronization information is used to correct the observed arrival times. This information is supplied from the network as assistance data and contains the time difference between carriers from different base stations. Finally, the arrival times are used to triangulate the position of the terminal.

E-OTD allows two operational modes to be used: hyperbolic and circular modes. Figure 2 illustrates these procedures.

The hyperbolic mode joins the temporal information of two base stations to trace a hyperbola. This hyperbola defines the area in which the user is expected to be. The intersection of two hyperbolas allows a two-dimensional (2D) position to be estimated. Two hyperbolas cross at two points, although one of

them can be discarded according to the cell information or the RTT. The circular mode, on the other hand, uses the propagation-time values to trace circles that define the distance from the mobile user to each base station. Accordingly, three circles must be drawn to perform 2D positioning. Note that the E-OTD technique may be implemented as a mobile-based or mobile-assisted technique, independently of the operational mode selected.

If the radio access network operates synchronously (i.e., carriers from different base stations are synchronized), the impact of this technique on the network is low, but if the base stations are not synchronized (e.g., as in GSM and FDD-UMTS networks), new components are needed to measure the synchronization between base stations. These elements are known as location measurement units (LMUs). A ratio of one LMU for every three to five base stations is needed (Soiman & Wheatly, 2002). The cost of such components represents a constraint in E-OTD implementation. Furthermore, as concerns traffic, two new limitations appear. The first is that the synchronization information needs to be periodically retrieved by the network component that is responsible for managing the positioning process. This slightly increases the traffic in the core network. The second is that, in mobile-based solutions, all the synchronization data must be delivered to the user terminal as assistance data. New methods are currently being developed to overcome this (Barcelo et al., 2004).

The fact that at least three base stations are needed to compute a 2D positioning (four in the case of 3D positioning) may be a problem in rural environments where the number of base stations within sight is low. E-OTD exhibits medium consistency since multipath may alter the propagation-time values and consequently the position estimation. Accuracy depends on several factors, such as the position of the base

stations, the number of base stations available for measurements, and errors in the synchronization estimation. Many LBS providers are implementing or developing products that use this technique to geolocate (Cambridge Positioning System, 2005; Openwave, 2005).

Observed Time Difference of Arrival (OTDOA)

OTDOA is a time-difference-of-arrival (TDOA) technique, which operates in a similar fashion to E-OTD: it determines the position by trilateration, either in the user terminal (mobile-based mode) or in the network (mobile-assisted mode). This technique was developed for use in WCDMA networks, especially in UMTS and CDMA2000 networks. The synchronization issues presented by E-OTD also apply to OTDOA. Therefore, if the transmitters in the UMTS radio access network (UTRAN) are not synchronized (e.g., in FDD mode), then the relative time difference between the base stations (node Bs) must be provided. The main difference between E-OTD and OTDOA is that node Bs may integrate the capability of measuring the synchronization between base stations.

OTDOA must address one limitation to which E-OTD is not subject: hearability. Hearability issues may occur when the user terminal is very close to the serving node B. In this case, the reception of signals from other node Bs in the same frequency might be blocked. Several proposals have aimed to improve the hearability of neighboring node Bs (Porcino, 2001; Ludden & Lopes, 2000; 3GPP TS 25.305, 2004). The most significant are presented as follows:

- **Idle Period Downlink (IPDL):** All base stations stop transmitting for short periods called idle periods. During these periods, mobile users can measure signals from

Table 5. Comparison of OTDOA modes (Cambridge Positioning System, 2002)

Parameter	OTDOA	IPDL-OTDOA	SWB OTDOA	A-GPS
Accuracy (67%)	50- 150 m.	30- 60 m.	15- 30 m.	10- 20 m.
Accuracy (95%)	Unable to locate	150 m	80 m	Unknown
Coverage failure	Exclusion zone at cell center: ~30%; fallback to Cell ID	Failures occur around periphery of cell: ~5%; no suitable fallback	Exclusion zone at cell center: ~5%; Cell ID suitable fallback	Indoors or when line of sight to the sky is obscured
UE complexity	Medium	High	Low	High
Operational complexity	Medium	High	Medium	Low
Multipath tolerance	Low	Low	Good	Dependent upon line of sight to satellites
Meets E-911 Phase 2 requirements	No	Probably	Yes	Yes
Impact on network capacity	Negligible	High	Low	Low (assistance data)

other base stations, thus mitigating the hearability problem. Idle periods are announced through the radio interface so that terminals are aware of the time periods in which they should make the OTDOA measurements. Minimal changes are required in the terminal. Since the solution is based on downlink, several mobile users can simultaneously benefit from it. However, the presence of idle periods directly affects the downlink throughput of the system, which was quantified as 0.3-6% of the capacity (Cambridge Positioning System, 2002).

- **Time-Aligned Idle Period Downlink (TA-IPDL):** This mode represents an extension of IPDL, in which the idle periods are time-aligned between node Bs. This alignment involves less interference in the OTDOA measurements at the user terminal and consequently less error. However, the implementation of this mode may

have a great impact on the network, since all stations should synchronize the idle periods.

- **Positioning Elements (IPDL-PE):** The terminal determines its position by measuring radio signals from a number of positioning elements (PEs). These PEs are placed at surveyed locations other than those of node B. This improves the accuracy of the location system, because the transmitters can be placed in the best place in terms of location error. For example, for base stations, aligned triangulation gives poorer results; the decision to place PEs could be based exclusively on location criteria, regardless of the node B emplacement. This method can be used as stand-alone (OTDOA-PE) or with the IPDL mode (IPDL OTDOA-PE).
- **Software Blanking (SWB):** This method uses signal processing to minimize the impact of hearability. The terminal makes

the OTDOA measurements and reports them to the serving MLC (SMLC), which applies algorithms to separate the signals coming from different node Bs. This mechanism allows legacy terminals to be used.

As in the case of E-OTD, this technique offers medium consistency. Table 5 displays the QoS achieved with several proposals and shows how the SWB mode achieves an accuracy figure which fulfills the FCC requirements for emergency services (E-911).

Global Positioning System (GPS)

GPS is a family of location techniques that use the NAVSTAR satellite constellation to geolocate mobile stations. Unlike TDOA techniques such as E-OTD and OTODA, GPS technology operates with a satellite constellation that is fully synchronized. Consequently, a minimum of three satellites are required to calculate a 3D position and two for a 2D position. However, this is only possible if the clock in the user equipment presents a similar accuracy to that of the clocks integrated in the satellites, which is costly and impractical in most scenarios. Therefore, three and four satellites are necessary for 2D and 3D positioning respectively—that is, one more measurement is used to allow the user clock to synchronize with the satellite clock.

GPS technology provides high accuracy and excellent spatial coverage, as well as good/excellent consistency. However, it has the following limitations:

- **Response Time or TTFF:** The user terminal has to wait until the satellite map has been completely received, unless the map is already available due to previous location requests. The delay introduced may be up to 12.5 minutes.

- **Non-Line-of-Sight Coverage:** GPS technology requires a direct line of sight between the terminal and satellites, and performs poorly in indoor or dense urban scenarios.

The operational mode known as WAG (Wireless Assisted GPS) or A-GPS (Assisted GPS) overcomes the aforementioned issues. The GPS technique is assisted by assistance information provided by the wireless network. Depending on the data transmitted by the network, we can classify GPS techniques as follows:

- **Differential GPS** uses the satellite information received at known positions to remove the inaccuracies introduced by the selective-availability mode of the GPS technology.
- **Indoor GPS** may be helped by additional information from the network, which allows weak signals to be decoded.
- **Assisted GPS** includes any assisted solution. A-GPS usually delivers ephemerides and almanac information (i.e., satellite map and orbit prediction) to the GPS receiver in order to speed up the TTFF, but other information could be sent in order to improve the performance. Differential and Indoor GPS can work as particular cases of A-GPS technology.

Angle/Direction of Arrival (AoA/DoA)

AoA/DoA measures the angle of arrival of the signal from the mobile user at several base stations. It is possible to geolocate a user by combining two or more angles of arrival. The main drawback of this location technique is that it involves installing complex arrays of antennas for estimating the direction of the signals generated in the user terminal and the dependence of the accuracy on the distance between the mobile station and the base station, and on

the number of antennas involved in the location process. Accordingly, the consistency offered by this technique can be considered to be fair.

Fingerprinting

This technique uses the fingerprint of the radio features to locate the user. The radio interface must be completely calibrated in all places where the location service is going to be provided. This characterization may involve several parameters, such as the signal power received, the multipath level (number of components, their strength and phase, etc.). This is different from NMR: in NMR, the radio path is modeled, while here a data sheet with fingerprints of all spots within the area to be covered is performed beforehand. The procedure for obtaining the position is as follows: the radio features of the whole area to be covered are calibrated and fingerprints are obtained. When the user terminal receives a location request, it takes a snapshot of its radio status. Then, the user terminal reports the snapshot to the SMLC. Finally, the SMLC looks up the radio status that best matches the snapshot from the user terminal in a database and assigns the position stored in the database to the user.

The accuracy, yield, and consistency of this technique are good/excellent. However, this technique presents drawbacks in terms of scalability and maintenance. The database must be frequently updated, and the larger the coverage area, the larger the database. This fact may affect response times depending on the database implementation. This proposal seems to suit the scenario of WLAN networks.

Hybrid Techniques

Hybrid technology involves any location technique that in some way combines information provided by several location techniques in order to reduce the limitations of each technique when it is used as a stand-alone: all of these improve accuracy, coverage, and consistency (Barcelo & Martin-Escalona, 2004). The following approaches may be taken:

- **Loose coupling** is based on performing several positioning processes using different location techniques, so that all the positioning processes can be combined and the QoS improved. OTDOA can, for instance, be combined with Cell ID. Thus, in the areas where hearability prevents

Figure 3. From left to right, non-synchronized and synchronized hybridization approaches (Barcelo & Martin-Escalona, 2004)

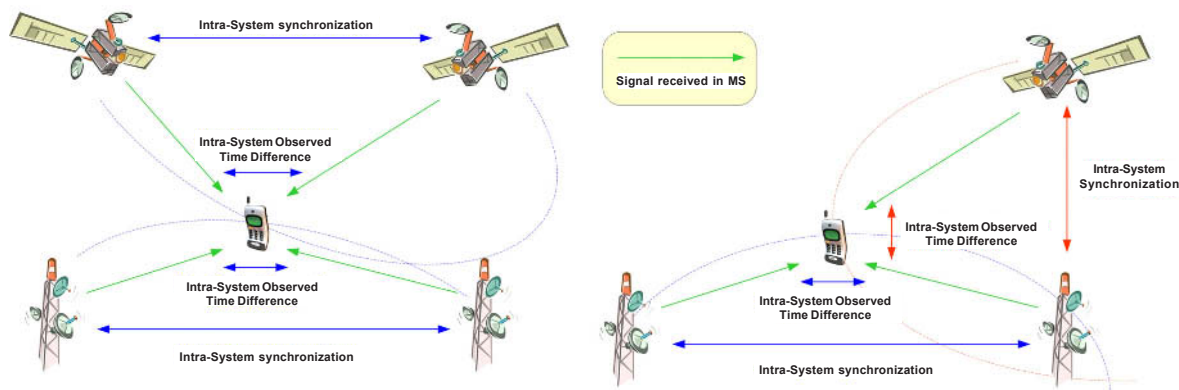


Table 6. Location techniques in GSM and UMTS networks

Technique	Network	Mode	Hybridization	Standardized
Cell ID	GSM, GPRS, UMTS	Network-based	With any location technique	Yes
Cell ID++	GSM, GPRS, UMTS	Network-based	With any location technique	Only Cell ID + RTT
U-TOA	GSM, GPRS	Network-based	With any location technique	Yes
E-OTD	GSM, GPRS	Mobile-assisted Mobile-based	With A-GPS and Cell ID	Yes
OTDOA	UMTS	Mobile-assisted Mobile-based	With A-GPS and Cell ID	Yes
IPDL- OTODA	UMTS	Mobile-assisted Mobile-based	With A-GPS and Cell ID	Yes
A-GPS	GSM, GPRS, UMTS	Mobile-assisted Mobile-based	With any location technique	Yes

OTDOA from being used, Cell ID is used as a backup technique.

- **Non-synchronized coupling** combines trilateration measurements from several location techniques (i.e., instead of combining positions). The signals used by one technique are not synchronized with the signals received by another technique (Rooney, Chippendale, Choony, Le Roux, & Honary, 2000). For example, if WAG and OTDOA are hybridized, the requirements for a 2D positioning are at least two node Bs and two satellites; as shown in Figure 3, two hyperbolas can be generated, one with the two OTDOA measurements and one from the two GPS measurements.
- In **synchronized coupling**, the signal transmitters are fully synchronized (Soliman et al., 2000). This reduces the number of measurements needed for geolocation. Figure 3 shows synchronized hybridization between OTDOA and WAG

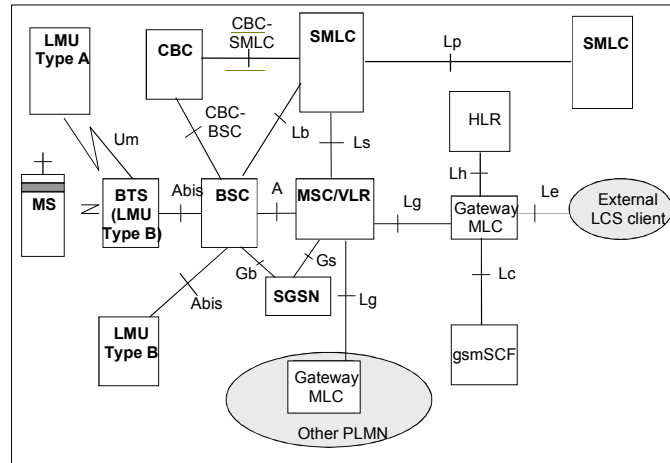
for the case of 2D positioning. Since all the measurements are synchronized, three signal transmitters can be used to perform the 2D positioning, whether they are node Bs or satellites.

LOCATION SYSTEM ARCHITECTURES

Today, most cellular networks allow for the provision of LBSs. Regulators have defined which location techniques can be used to carry out the positioning process. Table 6 gathers such information for GSM and UMTS networks.

The added value provided by location-based services led regulatory bodies to include their provision in the service definition for the main wireless networks. An example of this is the upgrade that ETSI made to the GSM network architecture to allow location-based services to be provided in 2G networks (3GPP TS 03.71,

Figure 4. Standard location system architecture for GSM/GPRS networks



2002): once location technology had evolved, ETSI decided to update the network definition to ensure network equipment compatibility in LBS provision. Nokia, Ericsson, and Motorola developed their own location solutions for GSM networks. The location architecture for this kind of network is displayed in Figure 4. The functionalities of the main components are explained as follows:

- **Location Client (LCS client):** A location client is any component inside or outside the operator's network that is able to generate location requests, which can then be carried out by the network.
- **GMLC (Gateway MLC):** A GMLC acts as a location gateway—that is, it is the interface used by the LCS client. A GSM/GPRS network may have several GMLCs, which perform the following tasks:
 - **Managing Location Requests:** To ensure interoperability between LCS clients and GMLCs, the OMA developed the Mobile Location Protocol (MLP) (Open Mobile Alliance [OMA], 2004). Once the location request has been received by the GMLC, it checks whether it has enough rights to request a position in that network. This verification involves consulting the HLR through the L_h interface.
 - **Managing Location Responses:** The GMLC is responsible for forwarding the position calculated in the network to the location client.
- **SMLC (Serving MLC):** A network can include several SMLCs to manage the location process—that is, it can provide all the facilities needed to run the location techniques and calculate the position. This involves the following tasks:
 - **Gathering the Location Requests from Several GMLCs:** The location requests are then distributed among several serving components.
 - **Managing the Positioning:** This facility involves several tasks. The SMLC can verify that the LCS client is allowed to request the customer's position and retrieve the assistance data if necessary. It also manages

the execution of the location technique, sending assistance data, and forwarding the request. The SMLC calculates the position if a network-based or mobile-assisted location technique is being run. Finally, the SMLC manages the result of the positioning process (i.e., checks for errors, verifies that QoS requirements are fulfilled, etc.).

- **Delivering the position** (or an error indication) to the GMLC for it to later forward to the LCS client.
- **Location Measurement Unit (LMU):** This element measures synchronization between the base stations. LMUs are only necessary if Time Difference of Arrival techniques, such as E-OTD, are used.

The approach followed by 3GPP in the definition of UMTS networks is different, since it includes LBSs as basic services in the UMTS network from the start. The location architecture proposed by 3GPP has grown out of the model presented by ETSI for GSM networks. The main differences between these two location architectures are in the GMLC facilities. The 3GPP proposal allows the GMLC to deliver location requests to both GSM/GPRS and UMTS networks. This means that if the user has a type-3 terminal (or even a type-2 terminal), the GMLC will have to decide which network is to manage the location request according to given criteria.

The architectures proposed by ETSI and 3GPP have been implemented by several network manufacturers. The proposal for location architecture made by Ericsson (Swedberg, 1999) includes two elements, the GMPC and SMPC, which implement the facilities assigned to the GMLC and SMLC respectively. Another example is the *mPosition* (Nokia, 2004), which is a solution proposed by Nokia, and is also based on the recommendations of ETSI and 3GPP.

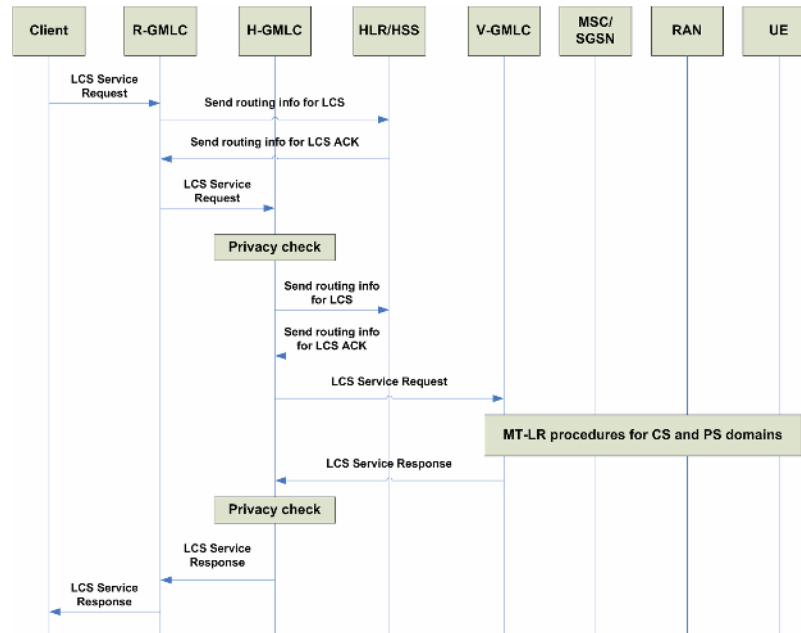
Location in cellular networks can be performed in three different ways according to the procedures stated by 3GPP and ETSI: Mobile-Terminating Location Request (MT-LR), Mobile-Originating Location Request (MO-LR), and Network-Induced Location Request (NI-LR). Mobile-Terminating Location Requests involve any location request in which an external LCS client asks the GMLC for the position of a network user. MO-LR procedure is run whenever the user wants to ask for his or her own position, either to use this information in an application running in the terminal or to deliver this information to an external LCS client. Finally, an NI-LR procedure is started by the network as a consequence of a user's status. The clearest example of this procedure is an emergency call: whenever a customer makes an emergency call, the network starts a location procedure (i.e., an NI-LR) in order to geolocate him or her and thus to provide the necessary information to an emergency center.

These three modes have a specific procedure depending on the domain the user terminal is working on: circuit-switched or packet-switched. The implementation of these procedures differs slightly between GPRS and UMTS networks. As an example, Figure 5 shows the common procedure followed in UMTS networks for MT-LR, either in circuit (CS) or packet domain (PS).

The steps followed in the MT-LR procedure are described below. More detailed information about all these procedures can be found in 3GPP TS 23.271 (2004).

1. The LCS client sends the location requests to the GMLC, which is known as R-GMLC (Requesting GMLC).
2. If the GMLC does not know the address of the user's Home-GMLC, it asks the HLR/HSS for this information. Otherwise, the procedure directly goes to Step 4.

Figure 5. Flow chart for MT-LR in UMTS networks (3GPP TS 23.271, 2004)



3. The HLR/HSS verifies that the R-GMLC is authorized to request the UE location information. Then, if this is the case, the HLR/HSS returns the H-GMLC address to the R-GMLC, along with several additional data.
4. If the R-GMLC is different from the H-GMLC, the R-GMLC sends the location request to the H-GMLC.
5. The H-GMLC verifies whether the R-GMLC is authorized to request UE location information. The H-GMLC then performs a privacy check on the basis of the UE user's privacy profile, which is stored in the H-GMLC, and the capabilities of the serving nodes (MSC/VLR and/or SGSN) if they are available.
6. The H-GMLC requests the IMSI for a particular MSISDN, the VMSC/MSC, or the SGSN address to the HLR/HSS, if any of these data are not known in the H-GMLC.
7. The HLR/HSS returns the requested information to the H-GMLC.
8. If Steps 6 and 7 were performed, the H-GMLC may perform a new privacy check. Besides this, if the Visited GMLC (V-GMLC) is different from the H-GMLC, it sends the location request to the V-GMLC. Otherwise, Steps 8 and 10 are omitted.
9. A circuit-switched or packet-switched location procedure is performed, according to the type of active session and the location technique.
10. The V-GMLC sends the location service response to the H-GMLC.
11. Once the location response is received by the H-GMLC, an additional privacy check may be carried out.
12. The H-GMLC delivers the location response to the R-GMLC. At this point, the H-GMLC may store some data in order to speed up future location requests.

13. The R-GMLC reports the location to the LCS client.

LOCATION MIDDLEWARE

Some years ago, the main issue confronted by location researchers was being able to provide the industry with accurate and consistent location techniques. As a result, a set of powerful location techniques came into being. Regulator bodies then started work on standardizing location in mobile networks. In fact, location services can currently be provided in widely deployed networks, such as GSM/GPRS or UMTS. Once the provision of location services was regulated in the most important mobile networks, location service providers realized that they had to develop services for specific networks, which entailed significant efforts in terms of implementation and maintenance. Providers thus came to demand a quick and easy method for developing LBSs, which was when location middleware came into being.

A location middleware is an entity placed between the LCS client and the mobile network which is responsible for isolating the LCS client from the technology used to get customers' positions. A location middleware can be logically structured in five components:

1. **LBS Manager:** This module is the core element of the middleware. It stores a list with all the implemented LBSs and manages the task of providing them. This component manages the message flow with the mobile network where the user is located. It also manages the location events that result from certain pre-programmed conditions and the location data.
2. **Navigation Information:** This module manages the navigation information, keeps these data updated, and matches the position obtained with navigation maps.

3. **APIs/SDKs:** This module provides interfaces that allow applications to use several facilities in order to exchange information with the middleware and thus make the development of LBSs easier.
4. **Network Drivers:** This component implements the drivers used to exchange information between the middleware and whichever mobile networks the customers may be using.
5. **Assistance Data Provider:** This module gathers all the assistance data requested by the location processes and is responsible for providing this information to the LBS manager whenever it asks for.

Nowadays, there are many location middleware available (SUN, 2005; Cell Point, 2005; Spanoudakis et al., 2003), and several of them are actively being used in running location platforms.

FUTURE TRENDS IN LOCATION

The use of location information entails benefits in many fields. Some of these are related to the optimal use of resources inside the network, and examples can include the development and implementation of location-aided handoff strategies (Goebbels, Siebert, Schinnenburg, & Lott, 2004) or internal middleware under the control of the network operator (Martin-Escalona & Barcelo, 2004). Those especially relevant become middleware. The functionality of these pieces of software could be extended in order to cover several issues that nowadays are not directly addressed. Some of them could be the impact of the user's mobility on the positioning QoS (e.g., the faster the user moves, the lower the accuracy and the shorter time the position can be cached), the influence of the network in the response time of location processes, and so forth.

Another application of location that has been taking on more importance as of late is geolocation in indoor environments, as these kinds of scenarios present special constraints if classical location techniques (e.g., A-GPS, OTDOA, etc.) are used. For example, indoor location might be used by emergency workers to locate partners in dangerous situations.

Location is also being applied to routing strategies in mobile ad-hoc networks (due to the high dynamism of these networks), and in flooding algorithms based on geographical zones (Seada & Helmy, 2004).

Another field where it is actively being working is in location-data management, in order to assure the privacy of user information.

Finally, it is worth mentioning that the EU is expected to update the E-112 regulations to set up a regulation framework as specific as the one proposed by the FCC in the United States. This would allow location techniques and infrastructures to evolve towards more competitive platforms, in addition to improving the capabilities offered to LBS designers.

REFERENCES

- 3GPP TS 03.71. (2002). *Functional stage 2 description of location services (LCS)*. ETSI. Retrieved from <http://pda.etsi.org/pda/queryform.asp>
- 3GPP TS 23.271. (2004). *Functional stage 2 descriptions of location services (LCS)*. R6. 3GPP. Retrieved from <http://www.3gpp.org/specs/specs.htm>
- 3GPP TS 25.305. (2004). *Stage 2 functional specification of user equipment (UE) positioning in UTRAN*. 3GPP. Retrieved from <http://www.3gpp.org/specs/specs.htm>
- Barcelo, F., & Martin-Escalona, I. (2004). Coverage of hybrid terrestrial-satellite location in mobile communications. *Proceedings of the 5th European Wireless Conference: Mobile and Wireless Systems Beyond 3G* (pp. 475-479).
- Barcelo, F., Martin-Escalona, I., Urruela, A., Riba, J., Manente, C., Gibeaux, M., & Deperini, F. (2004). *Method for localizing a mobile station in an unsynchronized mobile network* (PCT/FR04/08137). Patent pending.
- Cambridge Positioning System. (2002). Software blanking for OTDOA positioning. *Proceedings of TSG-RAN Meeting No. 16* (TSG RP-020372) (pp. 1-7).
- Cambridge Positioning Systems. (2005). Retrieved February 2005 from <http://www.cursor-system.com/cps/default.asp>
- Cell Point. (2005). *Cell Point location broker*. Retrieved February 2005 from <http://www.cellpoint.com/solutions.asp>
- Deirmentzoglou, S. (2004). Location-based services market perspective. *Proceedings of Mobile Location Workshop (MLW) '04*.
- European Commission Recommendation 2003/558/EC. (2003). *On the processing of caller location information in electronic communication networks for the purpose of location-enhanced emergency call services*. European Commission. Retrieved from http://europa.eu.int/eur-lex/pri/en/oj/dat/2003/l_189/l_18920030729en00490051.pdf
- Goebbels, S., Siebert, M., Schinnenburg, M., & Lott, M. (2004). Simulative evaluation of location-aided handover in wireless heterogeneous systems. *Proceedings of the 15th IEEE International Symposium on Personal, Indoor, and Mobile Radio Communications* (Vol. 2, pp. 1080-1084).
- Goodchild, M. F. (2001). *Final report of specialist meeting on location-based services*.

Santa Barbara, CA: Santa Barbara Center for Spatially Integrated Social Science and the University Consortium for Geographic Information Science.

Ludden, B., & Lopes, L. (2000). Cellular-based location technologies for UMTS: A comparison between IPDL and TA-IPDL. *Proceedings of the Vehicular Technology Conference* (Vol. 2, pp. 15-18).

Malenstein, J., Ludden, B., Pickford, A., Medland, J., Johnson, H., Brandon, F. et al. (2002). *Report on implementation issues related to access to location information by emergency services (E-112) in the European Union*. CGALIES.

Martin-Escalona, I., & Barcelo, F. (2004). Optimization of the cost of providing location services in mobile cellular networks. *Proceedings of the 15th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications* (Vol. 3, pp. 2076-2081).

Nokia. (2004). *Location aware, applications take off*. Finland: Nokia.

Open Mobile Alliance (OMA). (2004). *Mobile location protocol*. Retrieved from <http://www.openmobilealliance.org>

Openwave. (2005). *Location services*. Retrieved February 2005 from <http://www.openwave.com/us/>

Porcino, D. (2001). Performance of an OTDOA-IPDL positioning receiver for 3GPP-FDD mode. *Proceedings of the 2nd International Conference on 3G Mobile Communication Technologies* (Vol. 477, pp. 221-225).

Porcino, D. (2001). Standardization of location technologies. *Presentation at the Mobile Location Workshop (MLW) '01*.

Rooney, S., Chippendale, P., Choony, R., Le Roux, C., & Honary, B. (2000). Accurate

vehicular positioning using a DAB-GSM hybrid system. *Proceedings of the IEEE Vehicular Technology Conference* (pp. 97-101).

Seada, K., & Helmy, A. (2004). Efficient geocasting with perfect delivery in wireless networks. *Proceedings of the Wireless Communications and Networking Conference* (Vol. 4, pp. 2551-2556).

Soliman, S., Agashe, P., Fernandez, I., Vayanos, A., Gaal, P., & Oljaca, M. (2000). GpsOne™: A hybrid position location system. *Proceedings of the 6th IEEE International Symposium on Spread Spectrum Techniques and Applications* (Vol. 1, pp. 330-335).

Soliman, S. S., & Wheatley, C. E. (2002). Geolocation technologies and applications for third generation wireless. *Wireless Communications and Mobile Computing*, 2, 229-251.

Spanoudakis, M., Batistakis, A., Priggouris, I., Ioannidis, A., Hadjiefthymiades, S., & Merakos, L. (2003). Extensible platform for location based services provisioning. *Proceedings of the 4th International Conference on Web Information Systems Engineering Workshops* (pp. 72-79).

SUN. (2005). *Solution for location enablement*. Retrieved February 2005 from http://www.sun.com/solutions/documents/solution-sheets/TE_iforceloc_FF.xml

Swedberg, G. (1999). Ericsson's mobile positioning system. *Ericsson Review*, 4, 214-221.

TruePosition. (2004). *An examination of U-TDOA and other wireless location technologies: Their evolution and their impact on today's wireless market*. White paper. True Position.

Wilde, G. (2002). *Why are LB\$ a long time coming?* White paper. BWCS.

Chapter IV

Ontologies for Location-Based Services

Matthias Brantner

University of Mannheim, Germany

Sven Helmer

University of London, Birkbeck College, UK

Carl-Christian Kanne

University of Mannheim, Germany

Guido Moerkotte

University of Mannheim, Germany

ABSTRACT

A user of a mobile business application is usually not interested in technical data, but its meaning (which may also vary from user to user). We discuss how ontologies can help in translating this technical, location-based data (e.g. geographical coordinates) into semantic information. Taking a practical point of view, we first define typical requirements of location-based services, develop an ontology for locations, and show how this ontology can be integrated into existing technologies.

INTRODUCTION

One of the key advantages of mobile computing is the potential to personalize the services offered to a user based on his or her location. While the software industry has moved to-

wards service-oriented architectures (SOAs) in the last few years, this has mostly been done for non-mobile enterprise systems. In order to implement SOAs in a mobile context, we have to be able to handle location-based information effectively. Integrating different services on a

syntactical level—that is, exchanging geographical coordinates (e.g., GPS coordinates)—will not be enough. For users as well as applications, this form of location-based information is of little practical value. We need to attach meaning to the coordinates. However, this meaning also depends heavily on the user, as for example a tourist might be interested in the history of a building, a businessman in the companies located there, while someone delivering a package wants to know the address (street name and number). Additionally, the services themselves also need location-based information.

Here we discuss how ontologies can help solve these problems. When writing this chapter we chose to adopt a practical view. This is reflected by the different parts of the chapter: First, we give a brief introduction to ontologies and services (readers who are familiar with these topics can skip this). In the section on use cases, we define typical requirements of m-commerce applications with the help of concrete use cases. Based on these requirements we present an ontology for locations in the section on ontology. We explain the different elements of the ontology (written in OWL) and show how the previously defined requirements are met. We also demonstrate how this ontology can be integrated into existing technologies, taking as examples OWL-S and UDDI. Finally, a short summary concludes the chapter.

PRELIMINARIES

In this section we briefly define the terminology that is used in the following parts of the chapter. This mainly involves definitions for the terms ontologies and services.

For ontologies we include a rough overview of the Web Ontology Language (OWL) (World Wide Web Consortium, 2004b), while for services we will talk about service descriptors and

the interaction between service requesters and service providers.

Ontologies and OWL

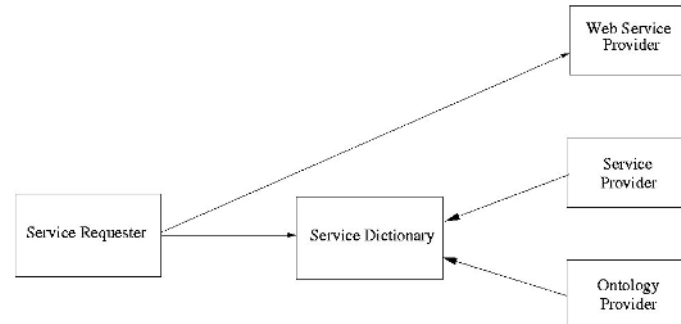
An *ontology* is an explicit specification of a conceptualization (Gruber, 1995). It specifies for a special domain the objects and relationships between the objects. In our context we use ontologies to describe services. Within these descriptions we have specialized ontologies that provide information on locations and regional availability of services.

OWL is a language to describe ontologies. The World Wide Web Consortium (W3C) developed OWL by revising the DAML-OIL Web Ontology Language. It defines semantic markup for Web resources and builds up on RDF (Resource Description Framework) (World Wide Web Consortium, 2004a). RDF is a general-purpose language for representing information on the Web, based on an XML syntax. OWL adds language primitives to support a richer expressiveness like cardinality restrictions, restrictions on the scope of properties, or on characteristics of properties (e.g., transitivity or uniqueness).

Different categories of requirements for OWL lead W3C to define three sublanguages of OWL (descending in power of expressiveness): OWL Full, OWL DL, and OWL Lite. For the purpose of describing our location ontology, OWL Lite suffices. In addition to being easier to understand, OWL Lite also has another advantage important in the context of large-scale mobile services. Due to its limited expressiveness, reasoning and searching can be implemented very efficiently.

An OWL document (describing an ontology) consists of an optional header plus any number of elements. The most important building blocks are classes, individuals, and properties:

Figure 1. Typical configuration



- Classes in OWL Lite define groups of individuals (instances of classes) that belong together. The set of individuals of a class is called an extension of that class. Classes can be organized in hierarchies using the `subClassOf` statement. All OWL classes are derived from the most general class `Thing`. There is also a class called `Nothing` that has no instances.
- Properties are used to model relationships between individuals or data types (using `ObjectProperty` or `DatatypeProperty`, respectively) and can also be organized in hierarchies (using `subPropertyOf`).

We can add additional information to properties and their values. Examples are `inverseOf`, `TransitiveProperty`, or `FunctionalProperty`. Functional Properties, for example, are used in our location ontology. They may have at most one value for each individual.

There is also the concept of restriction, which defines how properties can be used by individuals. Global restrictions hold for all individuals, while local restrictions refer to individuals of a certain class.

For a detailed description of OWL Features, we refer the reader to the specification of OWL (World Wide Web Consortium, 2004b).

Services and Service Description

We adopt the service definition from O’Sullivan, Edmond, and Hofstede (2002) where a service is an entity that performs an action on behalf of another entity. This is a very general definition, as it does not say who the entities are (e.g., humans, vending machines, or computers). We are mainly interested in electronic services that are requested or used via electronic media. The predominant communication media in this case is the Internet. An important subset of electronic services relying on the Internet are Web services. For the remainder of this chapter, when we talk about services we mean Web services (World Wide Web Consortium, 2002).

Figure 1 shows the typical configuration used by parties working together via a Web service. The service dictionary, often called service broker, contains descriptions of services. These descriptions come from service providers who want to offer a service. Additional information—like classification categories for products (e.g., UNSPSC), industry classifications (e.g., NAICS), or our location ontology—is supplied by ontology providers.

A service requester searches (and discovers) a service with the help of the descriptions in the service dictionary. The description of the service a requester receives contains all the

information needed to call the actual Web service at the provider.

REQUIREMENTS ANALYSIS

Before developing a location ontology, we need to identify requirements of services in the context of mobile applications. For this purpose, we formulate and analyze use cases that cover typical tasks that need to be supported by a software environment for mobile location-aware services.

We do not assume a particular architecture or a specific application domain. Instead, our analysis yields a checklist of requirements which can be used to simplify the requirements analysis of specific architectures and applications. However, the example configuration from Figure 1 can be used as an underlying “infrastructure” for the following Web service use cases.

There are two major classes of actors in a system for mobile services, *service requesters* and *service providers*. We devote a subsection to the use cases for each of these classes. We conclude the section by reviewing the requirements contained in the use cases.

Service Requesters

Example use cases for end users include:

- A user is riding in a train from Berlin to Frankfurt and wants to rent a car using his mobile device. The car rental company or at least their cars should be available near Frankfurt’s main station.
- A user is looking for a fast food restaurant near Frankfurt’s main station. He is aboard a train to Frankfurt’s main station and wants to use his mobile device.
- A user wants to have pizza delivered to his current location. Before doing so, he would

like to browse the pizza service’s menu using his mobile device.

Service Providers

Uses cases related to service providers include:

- A rental car company wants to register its rental stations.
- A rental car company wants to register its train-to-street service, which provides cars at every train station, even at train stations where the company does not have an office.
- A fast-food pizza franchise company, also offering delivery service, wants to register a new restaurant with its location and delivery area.

Requirements

In this subsection, we summarize the requirements implied by our use cases. This set of requirements will be used to assess the ontologies presented in the main body of our chapter.

- **RQ1. Searching for Services:** Service requesters want to easily search for services given an abstract formulation of their wishes. The result should be an accurate set of services (e.g., the names and addresses of fast-food restaurants).
- **RQ2. Using Services:** The service requesters may want to use some extended features, like viewing the pizza service’s menu or booking a car via a Web service. For this the service requester needs additional information (e.g., an URI of a Web service allowing him to book a car).
- **RQ3. Describe Services:** As service requesters want accurate results, Service Providers have to describe their service in detail with all necessary functional and

non-functional attributes. (For example, a pizza restaurant does not only want to publish its name and address, but also wants to provide an opportunity to reserve a table or to have a look at the menu.)

- **RQ4. Location Description:** The service requesters want to locate services using location information. This may not only refer to their current location, but to any target location. Consequently, services need some way to specify the regional availability of their service.
- **RQ5. Semantic Locations:** Purely geography-based location specifications are insufficient. The users of the system need to interact with the system using semantic concepts (e.g., city, street, and country names).
- **RQ6. Abstract Locations:** The ontology must be able to represent abstract locations which do not have a single physical manifestation (e.g., there is no such thing as *the* “train station”; there are hundreds and thousands of them).
- **RQ7. Inference Rules:** The ontology should be able to derive facts that have not been explicitly coded into the knowledge base. (For example, given the address of a restaurant, and the fact that the street it is located on ends at Frankfurt’s main station, it is possible to derive that the restaurant is near Frankfurt’s main station.)
- **RQ8. Simplicity:** Using the ontologies must be simple. The purpose of ontologies in mobile systems is to simplify their usage by bringing the user interface closer to the user’s view of the world. Mobile services are not only used or offered by trained specialists. There is no market for systems that require a traveler or a fast-food restaurant’s owner to understand descriptive logic formalisms.

- **RQ9. Extensibility:** The ontology is not fixed at system design time. Rather, providers add new types of services (e.g., the train-to-street service) and new types of locations (e.g., delivery areas).

ONTOLOGIES

In this section we present an ontology for locations and demonstrate how this ontology can be combined with existing technologies to offer location-based services. While doing so, we keep an eye on the requirements from the previous section and show how they are satisfied.

Ontologies for Locations

The use cases in this chapter make obvious that a mobile software environment needs a way to represent locations. In this section we discuss how the requirements summarized in this chapter influence the design of a location ontology. We also show how an ontology language such as OWL can be used to describe locations in a way that meets the requirements implied in the use cases.

Location Ontologies Based on Geographic Coordinates

A straightforward way of representing locations is to use coordinates based on some reference system. However, if this were our only location representation, users and service providers would have to specify service locations and queries using such coordinates. This would not only be tedious and error-prone, but may even be impossible, as the user may not have the precise coordinates of locations (e.g., the Frankfurt train station) that are relevant.

Hence, to realize a system that is friendly towards both service providers and users, we need to use a more expressive model, and only need geographic coordinates as a foundation tying the real world to the model.

We will not detail a coordinate-based location model here, as there are already standardized models and representation languages available (Open Geospatial Consortium, 2005b; ISO/TC 211 Working Groups, 2005). Our sample ontology below will reference individuals of such a model to describe the real-world position and extent of objects.

Location Ontologies Based on Graphs

As already mentioned, in a typical user-friendly mobile system, the user is rarely interested in the actual geographic coordinates of himself or the involved objects. Instead, he uses semantic concepts and symbolic names to describe his environment, such as “I am in Frankfurt” or “The restaurant is on Main Street” (RQ5). Moreover, the user does not think about the relations between the objects in a geographical way. In the restaurant example above, the precise topological connection between the street and the restaurant in the form of the common geographic boundary line is not relevant in the context of our use cases. Instead, mobile systems must be able to represent the fact that the restaurant is connected to the street, and that it is possible to pass between street and restaurant on foot.

A suitable model for such concepts is a graph, where the nodes designate object locations, and the edges describe the relationships between the objects.

In the following section, we present an example of such a graph-based ontology, and give examples of how to use individuals of the ontology classes to capture real-world information.

A Location Ontology

Our sample ontology will take the form of a graph with different types of edges for the different ways objects can relate to each other. We consider three types of edges, namely containment edges, specialization edges, and connection edges.

Since one of our requirements is simplicity (RQ8), we choose the easy-to-understand OWL Lite language to represent our ontology. One of the crucial points when designing OWL Lite ontologies is to decide for every concept whether it is best captured as an OWL class, an OWL property, or an individual. These three primitives of the language are interchangeable in OWL Full (e.g., in OWL Full, classes and properties can also have properties), but are disjoint concepts in OWL Lite. After presenting our OWL Lite ontology, we explain the associated design choices in Exhibit 1.

The aforementioned ontology is quite simple, but meets most of the requirements set forth in the use cases section, as explained in the following detailed description of its components:

- **Location:** The Location class is the core class of the ontology. It represents the nodes in our graph. Every conceptual location (RQ5) is an individual of the Location class.
- **Name:** The name property assigns a string-valued symbolic name to every location.
- **ContainIn:** The containIn property represents the containment edges in our graph. A location can be contained in another location, for example a street may be part of a city. Containment is transitive—a particular intersection of a street is also part of the city the street is in. Note that the graph induced by the containment edges need not be a tree, as for

Exhibit 1.

```

<owl:Class rdf:ID="Location"/>

<owl:DatatypeProperty rdf:ID="name">
  <rdf:type rdf:resource="#owl:FunctionalProperty"/>
  <rdfs:domain rdf:resource="#Location" />
  <rdfs:range rdf:resource="#xsd:string" />
</owl:DatatypeProperty>

<owl:ObjectProperty rdf:ID="containedIn">
  <rdf:type rdf:resource="#owl:TransitiveProperty"/>
  <rdfs:domain rdf:resource="#Location"/>
  <rdfs:range rdf:resource="#Location"/>
</owl:ObjectProperty>

<owl:Class rdf:ID="ConcreteLocation">
  <rdfs:subClassOf rdf:resource="#Location" />
</owl:Class>

<owl:Class rdf:ID="Locator"/>

<owl:ObjectProperty rdf:ID="locatedAt">
  <rdfs:domain rdf:resource="#ConcreteLocation"/>
  <rdfs:range rdf:resource="#Locator"/>
</owl:ObjectProperty>

<owl:Class rdf:ID="AbstractLocation">
  <rdfs:subClassOf rdf:resource="#Location" />
</owl:Class>

<owl:ObjectProperty rdf:ID="specializes">
  <rdf:type rdf:resource="#owl:TransitiveProperty"/>
  <rdfs:domain rdf:resource="#Location"/>
  <rdfs:range rdf:resource="#AbstractLocation"/>
</owl:ObjectProperty>

<owl:Class rdf:ID="TransportMethod">
  <rdfs:subClassOf rdf:resource="#AbstractLocation" />
</owl:Class>

<owl:Class rdf:ID="Connection"/>

<owl:ObjectProperty rdf:ID="originates">
  <rdf:type rdf:resource="#owl:FunctionalProperty"/>
  <rdfs:domain rdf:resource="#Connection"/>
  <rdfs:range rdf:resource="#Location"/>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="reaches">
  <rdf:type rdf:resource="#owl:FunctionalProperty"/>
  <rdfs:domain rdf:resource="#Connection"/>
  <rdfs:range rdf:resource="#Location"/>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="appliesTo">
  <rdfs:domain rdf:resource="#Connection"/>
  <rdfs:range rdf:resource="#TransportMethod"/>
</owl:ObjectProperty>

```

example a street that crosses a city border may be part of two cities.

- **ConcreteLocation:** Instances of the ConcreteLocation class are real-world objects that may have a physical manifestation, which is specified using a Locator.
- **Locator:** Locators describe geographic areas. A full formal specification of locators is beyond the scope of this chapter. Locators can be specified, for example,

using the GML language, and we refer the reader to the GML standard for more details (Open Geospatial Consortium, 2005a).

- **LocatedAt:** This property assigns a locator to a ConcreteLocation.
- **AbstractLocation:** An abstract location describes a class of locations, such as “train station” or “large city” (RQ6). Note that the actual abstract locations for a system are represented as *individuals* of AbstractLocation, not as *subclasses*.

The rationale behind this is to have a uniform design for service descriptions. The purpose of our location ontology is to allow the description of mobile services, in particular to describe those locations where the service is available or applicable. There are services that are more likely to be used, or even only applicable, in a certain *type* of location. For example, to activate a public transport timetable application is particularly useful if the user is at *any* bus station, not only for particular bus stations. To allow the designers of a mobile system to describe the availability or applicability of services in a uniform way, it is desirable to treat abstract locations in the same way as concrete locations. Another reason for choosing to represent abstract locations as individuals is explained under **Connection** as follows.

- **Specializes:** The specializes property describes both instantiation of an abstract location and subclass relationships between abstract locations. We model abstract locations not as classes, but as individuals. This makes abstract locations nodes in our location graph. Hence, we need another type of edge in our graph that describes the relationships between abstract locations, and between abstract locations and concrete locations. We real-

ize this edge type as a property between an abstract location and any other location.

- **TransportMethod:** A transport method is a subclass of abstract location that represents a way to move between locations.

Why does it make sense to treat transport methods as locations? As with abstract locations, the reason is uniformity for service descriptions. There are services that are applicable only when the user operates a vehicle, for example traffic congestion information. The specification of such a service is much easier if transport methods are represented similar to locations. Another reason is that we want to qualify connections between locations (see below under **appliesTo**).

- **Connection:** Instances of the connection class describe that two location individuals are adjacent. This is the final and most important kind of edge in our graph.

A straightforward choice would be to incorporate connectivity as a transitive property of locations. However, the semantics of such a binary relationship would be insufficient. For most mobile applications, it makes sense to qualify connection edges by a TransportMethod. For example, two streets in a pedestrian zone are connected for pedestrians, but not for car drivers. To capture this information would require n-ary properties which can link more than two individuals. However, OWL Lite only has binary properties. This is why we model connections not as properties, but as individuals of the Connection class. Each connection is a triple of a connection's origin, destination, and allowed method of transport. The three components of the triple are represented using the three functional properties **originates**, **reaches**, and **appliesTo** (see as follows).

izes, reaches, and appliesTo (see as follows).

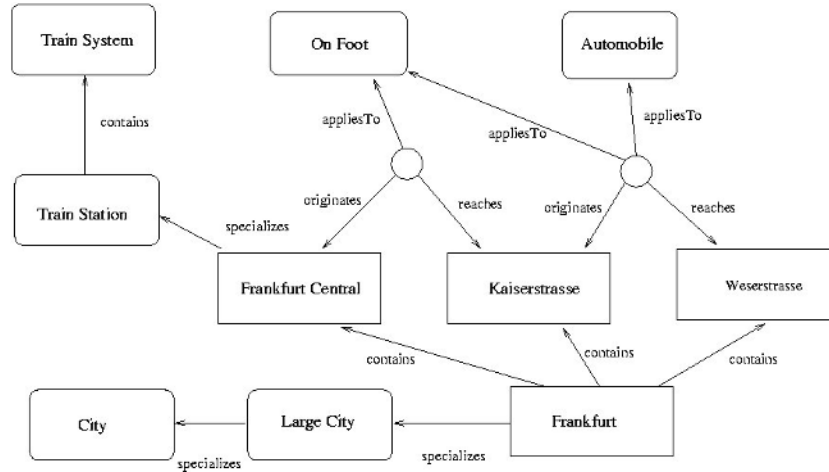
Note that connections are not reflexive. If there is a two-way connection between two locations, there must be one Connection individual for each direction. This is necessary to allow representation of directed connections, such as one-way roads or public transport lines that operate only in one direction.

- **Originates:** Represents the origin of a connection. Every connection has exactly one origin, making this property a functional property in OWL terms.
- **Reaches:** Represents the destination of a connection. As with origin, each connection

Exhibit 2.

```
<AbstractLocation rdf:ID="TrainSystem">
  <name>Train System</name>
</AbstractLocation>
<AbstractLocation rdf:ID="TrainStation">
  <name>Train Station</name>
  <containedIn rdf:resource="#TrainSystem"/>
</AbstractLocation>
<AbstractLocation rdf:ID="City">
  <name>City</name>
</AbstractLocation>
<AbstractLocation rdf:ID="LargeCity">
  <name>Large City</name>
  <specializes rdf:resource="#City"/>
</AbstractLocation>
<ConcreteLocation rdf:ID="FRA">
  <name>Frankfurt</name>
  <specializes rdf:resource="#LargeCity"/>
</ConcreteLocation>
<ConcreteLocation rdf:ID="FRACentral">
  <name>Frankfurt Central</name>
  <specializes rdf:resource="#TrainStation"/>
  <containedIn rdf:resource="#FRA"/>
</ConcreteLocation>
<ConcreteLocation rdf:ID="KaiserStr">
  <name>Kaiserstrasse</name>
  <containedIn rdf:resource="#FRA"/>
</ConcreteLocation>
<ConcreteLocation rdf:ID="WeserStr">
  <name>Weserstrasse</name>
  <containedIn rdf:resource="#FRA"/>
</ConcreteLocation>
<TransportMethod rdf:ID="OnFoot">
  <name>On Foot</name>
</TransportMethod>
<TransportMethod rdf:ID="Automobile">
  <name>Automobile</name>
</TransportMethod>
<Connection rdf:ID="C1">
  <originates rdf:resource="#FRACentral"/>
  <reaches rdf:resource="#KaiserStr"/>
  <appliesTo rdf:resource="#OnFoot"/>
</Connection>
<Connection rdf:ID="C2">
  <originates rdf:resource="#KaiserStr"/>
  <reaches rdf:resource="#WeserStr"/>
  <appliesTo rdf:resource="#Automobile"/>
  <appliesTo rdf:resource="#OnFoot"/>
</Connection>
```

Figure 2. Typical configuration



tion has exactly one destination, so reaches is a functional property.

- **AppliesTo:** Describes the allowed method of transport for a Connection individual. Here we have another motivation of our design choice to model abstract locations, in particular transport methods, as individuals: in OWL Lite only individuals can be connected using object properties—if transport methods were OWL classes, we could not associate them with connections.

Note that *appliesTo* is not a functional property, as there can be more than one transport method by which two locations are connected. However, to avoid a large number of explicit *appliesTo* relationships, all applicable transport methods that have a common ancestor with respect to the *specializes* property can be specified using just the ancestor individual. For example, if one street can be reached from another by both “car” and “bus,” a connection individual that *appliesTo* “automobile” is sufficient.

Based on the class and property definitions given above, the ontology can be used in a

concrete mobile software environment by populating it with individuals. We give two examples next.

Example 1

Figure 2 shows a few sample individuals based on the ontology components described above, both as a graph and as an OWL document consisting solely of individuals. To keep things simple, none of the individuals is associated with geographic coordinates.

In the graph, ConcreteLocations are rectangular boxes, AbstractLocations are boxes with rounded corners, Connections are small circles, and object properties are shown as arrows.

The graph representation makes it easy to see how facts can be inferred from the explicit knowledge (RQ7). For example, the fact that the Kaiserstrasse can be reached by pedestrians from a train station is not encoded explicitly. By searching for a path between Kaiserstrasse and Train Station in our graph, however, the fact can be inferred by traversing specialization and connection edges.

Example 2

Existing location ontologies can easily be integrated into the one given above. Consider the

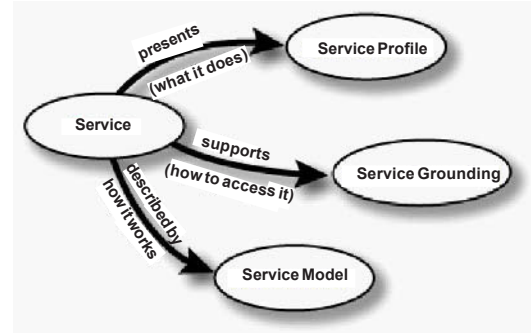
ISO3166 standard for country and province names (International Organization for Standardization, 1999) and the UN/LOCODE Code for Trade and Transport Locations (United Nations Economic Commission for Europe (UNECE), 2005). Basically, ISO3166 is a list of all the countries in the world and their subdivision into provinces, while UN/LOCODE lists important cities and ports, associating them with ISO3166 provinces. Together, these can be viewed as a very simple graph-based location ontology which can be expressed as individuals of the classes defined above. The countries, provinces, and cities are *ConcreteLocations*, and the containment relationship is described using the *containedIn* property. UN/LOCODE also gives geographic coordinates for the cities in form of points expressed using longitude and latitude coordinates. Each such longitude/latitude pair could be considered a *Locator* individual in our ontology.

Ontologies for Services

A location ontology cannot be used in an isolated way, but has to be embedded into a service. Important in this context is the service description: before a service can be used, it must be found; in order to be found, it has to be described accurately. When searching for services, their descriptions need to be comparable. This in turn requires a standardized, formal way to describe services. Current ontology-based approaches for service descriptions are, for example, UDDI, WSDL, and OWL-S (CORBA Services together with the CORBA Naming and Trading Service is an example for a non-ontology-based approach).

Our location ontology needs to be integrated properly into a service description to unfold its full potential. In the remainder of this section, we describe how this integration into OWL-S

Figure 3. OWL-S top level service ontology



and UDDI (which are the most important ontology-based service description technologies for Web services) allows us to offer services that satisfy the requirements derived at the beginning of this chapter. Both, OWL-S and UDDI offer extensibility of descriptions (RQ9) and enable service providers to describe their individual service with all functional and non-functional properties (RQ3).

OWL-S

OWL-S (David Martin, 2005)—formerly DAML-S (Ankolekar et al., 2001)—is an OWL-based Web service ontology. OWL-S allows Web service providers to describe any kind of Web service with the help of the following features:

- **Automatic Web Service Discovery:** This feature offers functionality to search for a specific service (RQ1).
- **Automatic Web Service Invocation:** After having discovered a specific service, the Web service invocation allows using the service and executing different actions (RQ2).
- **Automatic Web Service Composition and Interoperation:** This feature allows different Web services to work together;

for example, a car rental company could rely on another Web service to immediately process the payment.

- **Automatic Web Service Execution Monitoring:** This feature is not yet included in OWL-S. When eventually finished, it will allow services to signal their actual state and (if required) send notifications about, for example, state changes.

OWL-S defines an upper ontology for services, shown in Figure 3 (David Martin, 2005). The central class *Service* references three classes, each containing information to implement the features mentioned above. We now take a closer look at these three classes:

- **Service Profile:** The service profile provides a way to present the services offered by the service providers. Thus it acts as a superclass for every description of the service and manages three basic types of information:
- **Provider Information:** The provider information contains the attributes *serviceName*, *textDescription*, and *contactInformation* in a human-readable format.
- **Functionality Description:** The functionality description is a description of the functional properties (e.g., the transformation offered by the service). The transformation process (input to output) can be described using the Process ontology within the service model (see as follows).
- **Profile Attributes:** The profile attributes describe a lot of features that specify other, non-functional characteristics of the service. In OWL-S they explicitly contain the following attributes:
 - The Category of the service (e.g., an ontology from UNSPSC).
 - An extensible list of service param-

eters that further describe the properties of a service.

- **Service Model:** The service model within OWL-S describes how a service operates, which can be seen as a process. OWL-S defines a subclass called *process model*. The process model ontology allows modeling of an atomic process, a simple process, or a composite process. Both atomic and simple processes are single-step processes. The atomic process is directly invocable, whereas the simple process has no grounding and is abstract. A composite process is—as its name says—decomposable into other processes.

The process class is the superclass of the atomic, simple, and composite process classes. It contains various attributes to describe the process. These parameters include specifications of the in- and output, preconditions, and effects. These attributes can be used in the functionality description of the service profile and thus be helpful during the search for a service.

- **Service Grounding:** The service grounding specifies in a concrete way how to access the service. It describes how to map the abstract specifications of the service profile and the service model to communication primitives. However, it does not explicitly specify how to describe messages. This is done rather implicitly by the in- and output properties, using the well-specified Web Services Description Language (WSDL) (World Wide Web Consortium, 2001). (WSDL is an XML schema-based format that defines interfaces of Web services. Due to space constraints and the focus of the chapter on ontologies, we do not go into detail here.)
- **Integration of the Location Ontology:** As a location or a location-dependent avail-

ability is a non-functional property of a service, it can be described within the service profile's attribute list of service parameters (covering requirement RQ4). A service parameter consists of the parameter name (serviceParameterName) and some pointer to the parameter value (sParameterValue). This parameter value can be our location ontology described in OWL.

UDDI

Universal Description, Discovery, and Integration of Web services (UDDI) (OASIS, 2004) is a service discovery mechanism for Web services. The UDDI registry provides a standards-based approach to discover a software service (RQ1), invoke that service (RQ2), and to manage meta-data about the service

UDDI consists of two main parts: service registration and discovery. Here we focus on the registration part, as we will integrate our ontology into this part. Service providers register their services in a registry (RQ3), and service requesters query this registry for services according to their needs and wishes.

The information stored in the UDDI registry can be divided into three categories:

1. **White Pages:** The White Pages contain the name, address, and contact information of the service provider.
2. **Yellow Pages:** The Yellow Pages describe the service's type of business, the location of the service, and products the service offers. These descriptions are according to categorization taxonomies, for example geographical location by ISO-3166 or industry types by NAICS.
3. **Green Pages:** The Green Pages define technical information about the businesses' service, for example a pointer to a WSDL

file describing the grounding of the service.

The data structures of UDDI are expressed using complex types in XML schemas and basically consist of five parts:

1. **BusinessEntity** is the top-level data structure for a business offering services.
2. **BusinessService** contains a name and a description of a service being offered.
3. **BindingTemplate** stores information about the service according to the information represented within the Green Pages.
4. **tModel** consists of an abstract specification of a collection of information describing the service. Its representation is based on ontologies. In particular the tModel makes it possible to describe services using different ontologies (RQ9).
5. **PublisherAssertion** allows relating businessEntities to each other.

Integration of the Location Ontology

A general approach to extend the tModel has been taken by Paolucci, Kawamura, Payne, and Sycara (2003) and Pokraev, Koolwaaij, and Wibbels (2003) integrating OWL-S ontologies into UDDI. In the same way, we can hook up our location ontology (or others, like NAICS, UNSPSC, and ISO-3166) to UDDI. This will improve service discovery in the mobile context using UDDI considerably.

Further Extension

In the previous sections we have shown how to integrate non-functional properties into service descriptions with the help of ontologies. (For

OWL-S this can be done with service parameters in the service profile; in UDDI additional ontologies can be integrated into the tModel.)

There is no need to stop with our location ontology. In order to improve the quality of matching a service with the needs of a user, we would like to describe services as accurately as possible. So we need to formalize other, non-functional properties and implement them using ontologies. A lot of work has been done by O'Sullivan et al. (2005, 2002) describing how to formalize non-functional properties like charging, payment, prices, and also local and temporal availability.

SUMMARY

In this chapter we showed how ontologies can improve the quality of location-based services. Before presenting an actual ontology, we carefully defined the requirements of m-commerce applications in terms of location-based services. Based on the insights gained during the requirements analysis, we developed a concrete ontology for locations. Very important here is the concept of abstract locations and their connections with each other, which simplifies the search process considerably. In a last step we demonstrated how our ontology can be integrated into existing technology for describing services (OWL-S, UDDI).

Currently, we are working on a full-fledged location-based ontology in the context of the SALSA project at the University of Mannheim. Within this project our ontology will be applied in the framework of a service-oriented architecture.

Summarizing, we state that modeling locations via ontologies is a very promising approach to realize location-based services. However, as m-commerce applications are not en-

tirely established yet, it is very difficult at the moment to estimate the full impact that ontologies will have in this area.

REFERENCES

- Ankolekar, A., Burstein, M. H., Hobbs, J. R., Lassila, O., Martin, D. L., McIlraith, et al. (2001). DAML-S: Semantic markup for Web services. In I. F. Cruz, S. Decker, J. Euzenat, & D. L. McGuinness (Eds.), *Proceedings of the Semantic Web Working Symposium* (pp. 411-430). Stanford, CT: Stanford University.
- David Martin, E. (2005). *Owl-S: Semantic markup for Web services*. Retrieved from <http://www.daml.org/services/owl-s/1.1/overview/>
- Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing? *International Journal of Human-Computer Studies*, 43(5-6), 907-928.
- International Organization for Standardization. (1999). *Codes for the representation of names of countries and their subdivisions (ISO 3166)*. Retrieved from <http://www.iso.org/iso/en/prods-services/iso3166ma/index.html>
- ISO/TC 211 Working Groups. (2005). *ISO/TC211 standard*. Retrieved from <http://www.isotc211.org/>
- OASIS. (2004). *UDDI executive white paper and presentation*. Retrieved from <http://uddi.org/pubs/uddi-exec-wp.pdf>
- Open Geospatial Consortium. (2005a). *The Geography Markup Language (GML)*. Retrieved from <http://www.opengis.net/gml/>
- Open Geospatial Consortium. (2005b). *The OpenGIS documents*. Retrieved from <http://www.opengeospatial.org/specs/>

- O'Sullivan, J., Edmond, D., & Hofstede, A. T. (2002). What's in a service? *Distributed Parallel Databases*, 12(2-3), 117-133.
- O'Sullivan, J., Edmond, D., & Hofstede, A. T. (2005). *Formal description of non-functional service properties*. Retrieved from <http://www.citi.qut.edu.au/about/researchpubs/technical/nonfunctional.jsp>
- Paolucci, M., Kawamura, T., Payne, T. R., & Sycara, K. P. (2002). Importing the Semantic Web in UDDI. In C. Bussler, R. Hull, S. A. McIlraith, M. E. Orlowska, B. Pernici, & J. Yang (Eds.), *Proceedings of Web Services, E-Business and Semantic Web Workshop* (pp. 225-236). Berlin: Springer-Verlag (LNCS 2512).
- Pokraev, S., Koolwaaij, J., & Wibbels, M. (2003). *Extending UDDI with context-aware features based*. Retrieved from <http://citeseer.ist.psu.edu/641339.html>
- United Nations Economic Commission for Europe (UNECE). (2005). *United Nations Code for Trade and Transport Locations (UN/LOCODE)*. Retrieved from <http://www.unece.org/cefact/locode/service/main.htm>
- World Wide Web Consortium. (2001). *Web Service Description Language (WSDL)*. Retrieved from <http://www.w3.org/2002/ws/desc/>
- World Wide Web Consortium. (2002). *Web services activity*. Retrieved from <http://www.w3.org/2002/ws/>
- World Wide Web Consortium. (2004a). *Resource Description Framework (RDF)*. Retrieved from <http://www.w3.org/RDF/>
- World Wide Web Consortium. (2004b). *Web Ontology Language (OWL)*. Retrieved from <http://www.w3.org/2004/OWL/>

Section II

Health

Chapter V

Relevance of Mobile Computing in the Field of Medicine

Henrique M. G. Martins

University of Cambridge, UK

Matthew R. Jones

University of Cambridge, UK

ABSTRACT

Mobile information and communication technologies (MICTs) are widely promoted as increasing the efficiency of work practices in many business sectors, including healthcare. There are numerous types of mobile computing devices available that provide users with capabilities that can be applied in a wide range of different work settings. Case studies of the use of different MICT devices by doctors in different hospital settings indicate that while some doctors easily adopt MICT devices and find them a helpful tool, others encounter problems with their usage and, in fact, a majority do not use MICTs at all. This chapter deals with identification of five factors influencing the uptake of MICTs in clinical work practices and proposes a framework for analysing their interactions with the aim of increasing its uptake in medicine.

BACKGROUND

Mobile information and communication technologies (MICTs) are widely promoted as increasing the efficiency of work practices in many business sectors, including healthcare. There are numerous types of *mobile computing devices* available that provide users with capabilities that can be applied in a wide range

of different *work settings*. Case studies of the use of different MICT devices by doctors in different hospital settings indicate that while some doctors easily adopt MICT devices and find them a helpful tool, others encounter problems with their usage and, in fact, a majority do not use MICTs at all. This chapter deals with identification of five factors influencing the uptake of MICTs in clinical work practices and

proposes a framework for analysing their interactions with the aim of increasing its uptake in medicine.

INTRODUCTION

In most business sectors, including healthcare, it is widely claimed that the use of mobile ICTs, either alone or in combination with existing desktop resources, has the potential to achieve significant increases in the efficiency of work practices (Kelly, 2001; Cox, 2002; Davis, 2002). In recent years, however, a growing variety of *mobile computing devices* have become available—including laptop personal computers (PCs), tablet PCs, handheld PCs/personal digital assistants (PDAs), and PDA-phones/smartphones—which differ significantly in terms of characteristics such as screen size, computing power, weight, or input mechanisms, some of which have been shown to have an influence on ease of use and usage patterns (Dryer, Eisbach, & Ark, 1999; Martins & Jones, 2005). The devices may also vary in whether or not they are connected wirelessly to an existing network. This may be significant since, although when unconnected they are able to provide mobile computing power and support asynchronous communication, for uses requiring synchronous communication or real-time collaboration, wireless capability needs to be in place.

It is not just devices that vary, but also the work settings in which they are used. In the healthcare sector, there are a variety of departments in which hospital doctors' work, often organised according to particular clinical specialities. These departments may be spatially contained (e.g., in wards or intensive care units), or clinicians in certain specialities such as genetics, metabolic conditions, or psychiatric support may work across a whole hospital.

Less frequently, hospital doctors may work outside the hospital (e.g., accompanying acutely ill patients in transit from one location to another). In addition to potentially working in different physical settings, hospital doctors—like most highly skilled professionals—engage in several different types of activities over the course of their working day. These spatial and temporal dimensions of the organisation of doctors' work practices have been shown to influence how they use pen-and-paper and desktop ICTs (Westbrook, Gosling, & Coiera, 2004; Martins, Nightingale, & Jones, 2005).

Achieving the expected benefits from the use of MICTs, therefore, depends not simply on the provision of MICTs per se, but upon the appropriate matching of device characteristics and work settings. This chapter reports on research on the relationship between different types of MICT devices (laptop PCs on a cart or trolley, standard desktop PCs mounted on a cart or trolley, tablet PCs, and handheld/PDAs) and different clinical work practice¹ situations, and how this affects doctors' usage of MICTs.

RESEARCHING MOBILE COMPUTING IN HEALTHCARE

Two broad approaches may be used to study why and how doctors use (or do not use) MICTs in their clinical work practices: surveys across a large number of sites, or detailed studies in particular settings. This chapter largely focuses on the latter approach, presenting case studies of MICT usage at hospitals with different MICT devices and clinical settings.

Data were collected through multiple methods including interviews with doctors and hospital IT staff, observation, questionnaires, and analysis of usage logs for specific systems.

Case Descriptions

Case A: Paediatric Intensive Care Unit (PICU) with Handheld Computers

The PICU at a leading UK hospital had 12 beds, all located in a single ward. The layout of the unit comprised an open-plan central area with six beds and a nursing station (with two desktop computer) and a number of individual patient rooms and doctors' offices (with three desktop computers). The distance from the nursing station to any bed was not more than 10-12 metres.

The unit was staffed by about 7-10 doctors working on a shift pattern. The senior doctors, who had their own offices in the unit, were relatively permanent, while the junior doctors shared a common doctors' office and rotated between different departments. As a result, a particular doctor might be away from the unit for periods as long as 3-4 weeks. The department also sent doctors to outside locations to assist in the transfer of acutely ill patients to the PICU.

Desktop computers had been in use in the unit for a number of years, providing access to some basic patient demographic information, and ordering and reporting laboratory results. The unit provided a handheld computer for collecting data on ward rounds, which could be synchronized with the desktop PC in the doctors' office to update the departments' database. The handheld was never used consistently except for a highly IT-savvy doctor. Instead doctors would record data on pieces of paper during the morning ward round and then key in the data on the desktop PC later in the day. The reasons for this appeared to mostly relate to problems with synchronising the handheld and the time required to use it during the ward round.

The unit had previously supplied handheld computers (without wireless connection)—providing drug and medical reference information

and medical calculation applications—to some of the doctors. Despite this and the head of department's enthusiasm for handheld devices, only half of the 12 doctors interviewed used handheld devices in their work. Most of the doctors used them predominantly for arranging schedules and dairy appointments, whilst discussing a particular patient, and less frequently during ward rounds. Those that did not use a handheld device argued that they did not do so because patients tended to be on the ward for long stays and they could remember patient data and could access online information in their nearby offices if necessary. Some doctors also argued that due to the small screen, the technique required to input and access information on the handheld was less usable, if not difficult. This, according to them, was a reason for not carrying digital patient information, which they would have otherwise found useful. Interestingly, three female doctors said that since they had no pockets in which to carry the device, they only used it in the doctors' office. More than one doctor commented on the inconsistent way that colleagues used the handhelds, which restricted the possibility of it substituting for existing paper forms such as the unit job² list or pieces of paper used at shift handovers.

Those doctors using the handhelds found them helpful for storing information about drugs and infusions and telephone lists, and three of them stored short medical reference notes they created. One of the doctors also commented that he would use his handheld differently when working alone to record his own jobs from the shared list and set an alarm to remind him to carry them out, especially at night. Another enthusiast was one of the doctors involved in the retrieval of acutely ill children and who used his handheld extensively for these activities. Having customized most data to his needs, he said that he felt nervous about working without his handheld.

CASE B: Emergency Department with Handheld Computers

The emergency department (ED) of this U.S. hospital was seeing an average of 65,000 patients yearly and was staffed with about 45 physicians (attending and residents) working three regular eight-hours shifts. The area was a large, open space with four bays and several individual patient rooms. There were about 44 desktops available, four of which were located inside the individual rooms. The area had been rebuilt less than two years before, and the decision to install desktops was based on a fear that mobile computers (tablet PCs or mounted PCs) would be stolen or damaged easily due to the high patient throughput and staff turnover.

The department had issued stand-alone handhelds to all residents and handhelds/PDAs were made available to all senior doctors to increase doctors' access to specific departmental information (clinical pathways, protocols) and as a tool for personal organization (for example, call schedules, contact information, procedural information). Out of the 30 doctors surveyed, all had a handheld and two-thirds used it more than once a day. These frequent users, predominantly 'junior' doctors, valued the devices, arguing that they made work more efficient, saved the doctor from having to remember formulae, and allowed easier access to drug and medical reference information. Some of the senior doctors said that they used the handhelds less frequently, as they were familiar with the commonly used drugs as well as frequent pathologies from long experience. Another reason for less frequent use was the fear devices could be lost in a busy environment and the fact that local practices often diverged from those suggested on handheld-based applications. It therefore made more sense to discuss cases with colleagues and obtain information from them. Only a few doctors reported using

the devices to access hospital documents/guidelines, and none for accessing patient data, the Internet, or to send/receive clinical-work-related e-mails. The main reasons for this were said to be that it was hard to access data on the handhelds, especially compared to the widely available desktops. Department protocols, although available on handhelds, were also normally accessed via desktop PCs because doctors found the large screens more convenient than the small handheld ones and they could simultaneously access online resources (for medical and drug reference as well as for calculations).

CASE C: Breast Unit with Tablet PCs

This UK hospital breast unit comprised a multidisciplinary team of surgeons, radiologists, and physicians. Apart from short inpatient stays for surgery, the majority of contact with patients was via outpatient appointments at the unit. There would be about 5-6 clinic sessions per week when about 20 patients would be seen. Three physicians, two or three surgeons, and two radiologists would typically staff the unit. On a first visit patients would normally be examined by a physician and see a radiologist. On the second visit, physicians and occasionally surgeons would discuss results and the treatment plan. The unit had seven consultation rooms which had no desks, a large reporting room with three desktop computers where all team members (doctors, nurses, care managers) congregated during outpatient sessions, a waiting area, and a few offices for senior doctors.

Nine months previously a new clinical information system specifically designed for the unit was introduced. It was accessible via desktops in doctors' offices and the reporting room. The unit had also installed a wireless network and made available four tablet PCs for usage during

sessions, with the intention that physicians especially would use them to record patient data on the first visit.

Observation during several outpatient sessions revealed that only the radiologists used the tablet PCs in the reporting room as stationary devices (permanently plugged to a power supply and with a mouse installed). Physicians would take a few sheets of paper into the consultation room and then enter their handwritten notes via the three desktop PCs available in the reporting rooms. Surgeons rarely used any computers, but would instead dictate their comments and provide these recordings to secretaries for entry into the application.

Amongst the reasons offered by the physicians for not using the tablet PCs were that they would still need to write certain details down on paper (e.g., the patient's past history), as the application covered only part of their data recording needs and patient records were still largely paper based. Another issue was that it was more efficient for them to walk the short distance from the consultation rooms to the reporting room and enter the data on a comfortable and familiar device, rather than to feel awkward using a tablet PC to input data in front of a patient. This was especially the case with the text-based data, required by the nature of the speciality and patient history, which was felt as particularly difficult to enter using the active pen onscreen mechanism only. In the consultation rooms some doctors felt that having nowhere to place the device besides the chairs or the examination bed created a risk of it being dropped. Lastly, one of the physicians stated that she had not been made aware she could use the devices, and she thought that since radiologists were using them in the reporting room, they were for their exclusive use. The radiologists, for their part, predominantly used the tablet PCs because they did not have enough PCs in the area of the communal room

where they worked. The tablets could also be brought closer to the X-ray viewing boxes. The senior radiologist was an enthusiastic user, who took the tablet PC home with him and had loaded personal work files and customized it to his own use. Occasionally he would also use it as a laptop in other unit locations.

CASE D: Renal Unit with Mounted PC and Tablet PC

The renal unit of this UK hospital had 64 beds. Most inpatients were located in three wards on one floor of the hospital building, although there were usually a further 6-10 patients located in other wards around the hospital. Some of the senior doctors' offices were some floors away from the ward, and the outpatient consultation areas were in another nearby building. The doctors' office (mostly used by less senior doctors) was located in the middle of the central ward and had three desktop PCs. The mounted PC on a trolley/cart would also usually be left here when not in use.

A wireless network had been installed in the wards that allowed access to the department's e-prescription/lab reporting/requesting application and to the British National Formulary (BNF) Web site. Tablet PCs were shared with nurses, although the two battery-chargers were located in the doctors' office. In addition to the doctors' office, desktop PCs were available in the consultant offices, outpatient consultation room, nursing stations, and some ward clerks' offices. These provided access to the departmental application as well as other hospital applications and full access to the Internet.

Two medical teams of about three to six doctors (comprising one or two consultants, one or two registrars, and one or two house officers) covered the wards each day. On most mornings team members would get together for a "consultant ward round," while on other days

these would occur without the consultant. These ward rounds would start on the unit wards and then proceed to other floors of the hospital where their inpatients might be located. No doctor was seen picking up a tablet PC to visit an individual patient, although some said that they did so occasionally. Doctors were also not seen to use tablet PCs or the mounted PC except for morning ward rounds. Teams would use either one tablet PC or the mounted PC for each ward round, never both together, and occasionally on ward rounds without a consultant, neither would be used. The mounted PC was never taken outside the unit, but the team had started to take the tablet PC when visiting patients at a ward located in another floor of the hospital when wireless coverage was extended there. There were also a number of patients that the team would visit during the round at other hospital locations where mobile computing and the department application could not be used.

All doctors liked accessing the department application on the mobile computers for the ward round, although the need to change the tablet PC frequently because the battery had died, or loss of wireless connection with the mounted PC in some of the most distant areas of the ward, made some doctors consider using nearby desktops. Convenience of access and having information readily available in one location (as opposed to team members needing to search for different pieces of paper) was said to improve decision making, especially compared with when teams visited patients outside the unit.

The hospital had no full electronic medical record (EMR), so teams would conduct their rounds with the files of patients' notes as well as mobile computers. This meant that, in addition to whichever mobile computer was in use, the team also needed a trolley to carry the sometimes voluminous notes. While with tablet

PC this was not a big issue, it made the team reluctant to use the mounted PC if there were less than two junior doctors in addition to the consultant. Doctors' opinions were divided as to what they considered to be the better device for ward rounds. The mounted PC was seen as more cumbersome and as slowing down the rhythm of the round, because it could not be easily brought to the patient's bedside (it either stayed in the corridor or away from the immediate bedside area). Although the big screen of the mounted PC was seen as an advantage for information sharing and discussion, the mobility of the tablet PC between users was regarded as able to compensate for this. Such sharing of information and its collective use was not always observed or commented as unproblematic. Retaining control over the mobile computer or using it to show data on display as a backup for stronger argumentation in discussions occasionally seemed to reveal and was described as attitudes of power exertion.

Case E: Medicine Department with Laptops on Trolleys and Handhelds/ PDAs

The medicine department of the Veterans Affairs Medical Center, Washington, DC, had five medical teams (each one composed of one attending, one resident, two interns, and one or two medical students) covering about 72 inpatient beds on two consecutive floors. On each floor the ward had a central area (with several desktop computers) from which six corridors of patient rooms and two or three doctors' rooms (with five desktop PCs) led off. Service stairs connected the floors, but the lift (elevator) was located away from the centre of the wards. Normally teams had patients on both floors, and doctors identified this layout as affecting their usage of the Wi-Fi-enabled laptops made available to them, as there were attached to a food

trolley and this apparatus was only movable as a whole. The laptops had been available for more than five years and provided full access to Windows software, two hospital systems (a full EMR and a drug administration system mostly used by the nurses), the Internet, and a resource-rich hospital intranet.

About one-quarter of the doctors claimed that they never used the mobile devices, and only about one-fifth of the doctors were high-frequency users, using the laptops more than five times a day for clinical work. All doctors used the laptops to check patient data, but only two for medical calculations and one to send clinical-related e-mails. No doctor reported using the laptops for checking hospital guidelines.

About four senior doctors had recently been given a PDA to try accessing the EMR wirelessly, although the handheld application was not as comprehensive and could not be used, for example, to prescribe or to retrieve certain patient data. Only one of the four senior doctors occasionally tried to use her PDA regularly.

Doctors only used the laptops during team ward rounds and did not use them when visiting patient's rooms during the course of the day. A number of reasons were given for this, such as fear of losing/damaging the device, and that it was cumbersome to carry just for a quick visit to a patient room, especially as such visits often involved walking from one end of the department to the other or between floors. A number of doctors also said that they did not think they were allowed to use the device for tasks other than ward rounds.

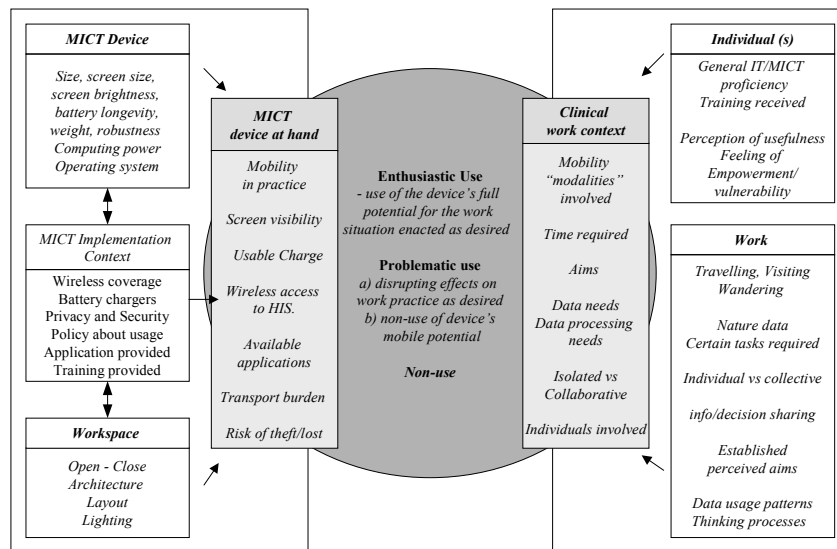
Three out of five medical teams in the department never used the laptops, while the others used them regularly (albeit only for ward rounds). This appeared to be related to differences in ward rounds practices, different perceptions of the function of ward rounds, and whether this could be supported by MICT. For

example, while one of the teams fostered the use of the laptops to access the Internet to look at online resources for retrieving pertinent medical information during ward rounds, others would not favour this although they might use them extensively to access the EMR. Individual doctors' attitudes towards technology, but also regarding the importance of rich and timely data in clinical decision making, the effectiveness of existing practices, and their willingness to change practices, also appeared to play a role. Strong social influences such as the power relationships between grades of doctors, existing work routines, and team dynamics were also associated with differences in use of mobile devices, as were departmental policy not clarifying that devices could be shared amongst teams on different floors.

CASE F: Clinical Genetics Specialist with Handhelds

In this large New York State hospital, there was just one doctor providing clinical genetics services, supported by a nurse and some genetic counsellors. The doctor's main work involved diagnosis of outpatients, who he would then follow during their subsequent medical history, but he also often visited inpatients around the hospital. He had developed his own patient database, accessible via his handheld, to which colleagues contributed via desktop PCs. As he acknowledged difficulties with entering so much most data via his handheld, he included input strategies in the application design, like extensive drop-down menus and the option of entering full-customized sentences that he used more frequently. In addition he could always access patient data even though he was often away from his office with no access to a desktop PC or connection to the database in his department. He often received phone calls about patient care either in the hospital or at

Figure 1. Schematic representation of the relationship between a particular MICT device at hand and a specific clinical work context



home. In both situations access to patient data and records of previous decisions was considered helpful, as was being able to record data immediately rather than when he next was in his office. The information on his handheld was shared with his secretary, who served as a "fixed" contact point. In addition to potential efficiency gains, he saw the possibility of accessing more complete information as a quality improvement, but also as a form of personal style regarding patients and families. The main reason, however, that made him keep using the handheld was for scheduled visits to inpatients around the hospital, where he regularly uses it by the bedside and then uses an infrared-enabled printer to produce a customized report to add to the hospital's paper record.

DISCUSSION

From these brief case descriptions, five themes may be identified that may be helpful in understanding the use (and non-use) of different

MICT devices in medical work (the relationship between the categories and themes is shown in Figure 1).

MICT Devices Characteristics

MICT devices differ in their technical characteristics, and these alone can influence use. Screen size is perhaps the most obvious example of this, affecting viewing but also data entry on handheld and tablet computers. Where handhelds were involved, (small) screen size was shown to limit doctors' willingness to input large amounts of data (Case A) or retrieve information (Cases B and E). With tablet PCs (Cases C and D), this was no longer referred to as an issue for data retrieval, but when these were contrasted with larger screen alternatives (desktop or mounted PCs), tablet PC screens were considered less suitable (Case C) and a deterrent to information sharing and collaborative working (Case D).

Another significant characteristic is the size of the device itself, which has implications for

the devices' *mobility in practice*. Handhelds are the most wearable MICT devices, but are still considered too big if the user does not have a suitable pocket to carry them in (Case A). When it comes to tablet PCs and PCs (or laptops) mounted on trolleys, size is again a distinguishing factor, as Cases D and E illustrate. Users had problems with mounted PCs/laptops on trolleys when their work involved a lot of walking. In Case D where a tablet PC alternative existed, it was often chosen due to size alone, and in Case E, it was clear that while the handheld would always be carried by the senior doctor into the patient's room, this was not the case with the laptop on the trolley. Size was also related to concerns about theft and damage, as was shown in Case B where the head of department provided doctors with handhelds because they could be personally owned and carried by professionals themselves.

Other characteristics such as screen brightness, battery longevity, weight, robustness, processing power, and operating system also influenced usage, both directly and in combination with other factors. This is in agreement with experimental studies (Dryer et al., 1999) that have shown that certain MICT device characteristics influence not just usage behaviours, but also users' attitudes towards the devices and towards other people using them.

Workplace Characteristics

As the cases illustrate, there is considerable variation in the characteristics of workplaces within hospitals. For example, while the size of some, such as the intensive care unit in Case A or the outpatient unit in Case C, was quite restricted and access tightly controlled, others such as the Accident and Emergency Department in Case B covered a larger area and were accessible to the general public. This had an impact on users' perceptions of the risk of theft or loss of devices. Similarly, the particular

layout of some wards and the location of desktop PCs (Cases D and E) may create incentives to use mobile devices, while widely (Case B) or readily (Case C) available desktop PCs may discourage use of mobile devices.

Again, while some of these characteristics like physical layout or access to the settings may directly influence usage, others, such as the architecture of the building and even lighting conditions, may significantly affect MICT use when combined with certain device characteristics by influencing the convenience and comfort with which devices can be used (Figure 1).

MICT Implementation Context

In addition to the technical characteristics of the MICT devices and those of the workplace per se, adoption may be influenced by an organization's decisions about how to deploy mobile technology and for what purposes. These may relate closely with hardware, like the extent and reliability of wireless coverage (Case D) provided or the battery charging options. In Case E, the decision to use the standard laptop battery only (to reduce the weight of the trolley) meant that batteries would sometimes die, even when fully charged at the beginning of a ward round, if the device was used extensively or the round lasted longer than usual. Another type of decision related to policy measures to ensure confidentiality, privacy, and security (including theft of the devices themselves) which restricted the flexibility with which mobile devices could be used. A prime example of this was the laptops in Case E that were fixed to trolleys to prevent thefts. In the same case, the lack of policy on sharing of devices meant that teams did not use available devices when visiting other wards.

Another significant implementation issue concerns the applications and level of Internet access made available on the mobile devices.

Applications obviously determine the particular data available and also how this is retrieved and manipulated. Some applications can overcome device limitations to a certain degree, as it was with the drop-down menus in Case F that compensated for the small handheld screen size. Others, however, may also highlight device limitations. In Case C, for example, the application required large text data fields to be entered, which was felt to be incompatible with data entry on tablet PCs, especially in full view of patients in the consultation rooms. Hardware-related training has been widely identified as a factor in the use of desktop technologies (Riley, Lorenzi, & Dewan, 2002), and it seems even more relevant with MICTs due to the new skills involved in data input on some devices and also their diversity.

Characteristics of the MICT Device at Hand

By *MICT device at hand*, we mean the mobile computing resource as a whole—hardware, software, and usage policies—as it presents itself to a particular context. For example, screen brightness in itself is not enough to ensure screen visibility; it also depends on the lighting conditions in a particular setting. *Mobility in practice* may also be seen as another instance of the contingent nature of the *device at hand*. Thus, while mounted PCs (Case D) or laptops on trolleys (Case E) were generally mobile when used on a single floor of the hospital, the exact same devices could not be used with some patients because of architectural barriers (stairs, narrow doorways).

Individual(s') Characteristics

Like MICT devices, individual doctors are not all the same. For example, age, seniority, and whether they are a surgeon or a physician may

have an influence on users' attitudes to, and usage of, MICT devices (McLeod, Ebbert, & Lymp, 2003; Barret, Strayer, & Schubart, 2004). Most authors (Brenda & Gadd, 2001; McLeod et al., 2003), however, have suggested that such differences, for example with grade, are probably linked to work specificities and the roles enacted, rather than individual dispositions. The cases suggest, however, that there are a few individual characteristics that may directly influence usage in terms of *how* (rather than *what*) clinical work practices are carried out. For example individual doctors undertaking retrievals in Case A or working in Accident and Emergency in Case B experienced the same *MICT device at hand* in a consistent *work context*, but exhibited different usage patterns.

Doctors' general IT proficiency (keyboard skills, knowledge of applications, and ability to troubleshoot problems), as well as their familiarity with aspects particular to using certain mobile devices (e.g., using *graffiti* to input data on a handheld/PDA, or using active pens on a tablet PC), may also influence not only the adoption but also the ease with which devices are used in work situations. This was illustrated in Case A, where most doctors had difficulties with updating the database; only one, recognized to be more enthusiastic and IT proficient than the norm, was able to do this easily. Personal perceptions about the value of MICT devices compared to existing tools, such as paper or desktop PCs, were also found to vary significantly. For example, some doctors in training in Cases A and B valued the reminders provided by the MICT (with information about drugs, medical references, to-do lists), while others preferred to rely on their memory or notes on pieces of paper, especially where desktop terminals were nearby (Cases A and C).

Individual perceptions about their role and how this should be carried out also varied. In

Case F, for example, the doctor's desire to be able to access every detail of his patients at all times meant that his handheld was seen not only as a useful tool, but one that actually improved efficiency and quality of care. Such perceptions could sometimes be shared across groups of doctors as was shown in Case E, where some teams were enthusiastic users, while others considered that the devices offered little or no advantage to their work and could even be a distraction.

A slightly different theme is that of doctors' feelings of empowerment, or conversely of vulnerability, when using mobile technologies. The concerns of physicians in Case C about using MICTs in front of patients may be contrasted with the positive feelings created by the personalized use of handhelds by the genetics specialist in Case F or the doctor involved in retrieval in Case A. There may also be a collective element to such feelings, as illustrated by the different responses of senior and junior doctors to the use of MICT during ward rounds in Case D.

Work Characteristics

Hospital doctors can potentially use MICT devices in a variety of work situations that vary along a number of dimensions over which they have more or less control. Work may vary, for example, in terms of their *modality of mobility*. Kristoffersen and Ljungberg (2000) distinguish three such "modalities"—travelling, visiting, and wandering—each of which has distinct characteristics: *travelling* is the process of going from one place to another, [often] in a vehicle; *visiting* is spending time in one place for a prolonged period before moving on to another place; and *wandering* is extensive local mobility in a bounded area.

Patient retrieval may be seen as an instance of travelling mobility (Case A), ward rounds as

an instance of visiting, and junior doctors' response to requests at other times as an instance of wandering (Cases D and E).

These modalities are not necessarily mutually exclusive and there may be differences within them, but they are seen as affecting the suitability of certain types of MICT devices. Another aspect of mobility in work practices is highlighted by Luff and Heath (1998) who discuss *micro-mobility* as "the way in which the artefact may be mobilised and manipulated for various purposes around a relatively circumscribed, or 'at hand' domain." For example, a piece of paper may be easily passed from one person to another or read by more than one person at a time. This was illustrated in Case D, where the tablet PCs were handed around among the team at the patient's bedside, but the laptops on trolleys in Case E could not be so easily shared.

Work may also vary with respect to the *level of mobility* involved—predominantly stationary or predominantly mobile—and this seems to affect users' assessments of the value of mobile devices. This is clearest where doctors can "wear" handheld devices (Cases A, B, and E), but for devices like tablet PCs and especially mounted PCs, usage was low in both predominantly stationary activities (e.g., work in the doctors' office with occasional visits to patient rooms) and in highly mobile situations (for example, visiting patients on different floors of the hospital). It was in situations where there was an intermediate degree of mobility with significant periods of static use such as ward rounds (Cases D and E) that doctors seem to prefer these types of MICTs.

The individual and collective nature of the work practice seemed to influence usage of MICT devices. For example, handhelds were mostly considered suitable only for individual use due to small screen size and difficulties of communicating between different models. The

team-based nature of some aspects of doctors' work, on the other hand, meant that the social influence of colleagues, especially those of senior grades, could encourage or discourage MICT use.

Perceptions about the objectives of MICT use clearly influenced how users individually (Case F) and collectively adopted them. In Case E, for example, three teams used the wireless laptops to access the EMR, but only one considered and encouraged their use to access online resources as well. Moreover, while the availability of nearby desktop PCs influences perceptions of the need for mobile access to data, this also appeared to depend on the type of data handled and the data analysis carried out. The genetics specialist in Case F, for example, felt that he needed a mobile device because he could be called on to make decisions about patients, based on complex data, at almost any time. Similarly, some senior doctors in Case E felt that the use of MICT devices on ward rounds was valuable to enable them to undertake more sophisticated analyses of patterns in patient data, rather than relying on the memory of the junior doctor presenting the case.

A Framework for Analysing MICT Use

From the above discussion we may identify two broad groups of influences on MICT use in clinical settings, one of which relates primarily to the characteristics of the device, earlier described as the *MICT device at hand*, and the other relating to the particular work practices in which they are employed, which we will describe as the *clinical work context*. This is illustrated in Figure 1.

The cases do not provide sufficient evidence to suggest what particular combination of these influences will lead to enthusiastic use

of MICTs, to problematic use of MICTs, or to non-use of MICTs, nor is it claimed that all possible influences have necessarily been identified. Figure 1 is presented, however, as an aid to conceptualisation of the types of influences that may need to be considered in understanding MICT use in clinical work practices, both at the macro level, in terms of addressing both technical and work practice issues, but also as an indication of potential influences (some of which appear to be distinctive to MICTs) within these two areas. The static representation of Figure 1 should also not obscure the dynamics of the balance of these influences in any setting—that is, a change in only one of the characteristics (e.g., loss of battery power or a new senior team member) may be sufficient to alter usage of MICT devices.

CONCLUSION AND FUTURE DIRECTIONS

In this chapter we have presented reports of MICT use in different hospital settings in four hospitals in two countries. Our own results in other countries and clinical settings suggest that it may be more broadly applicable, although it would be valuable if this could be confirmed by other studies. Perhaps the main contribution of this work is in drawing attention to the interaction between the technical and social influences on the use of a particular *MICT device at hand* in a specific *clinical work context*. A similar interplay may also be applicable in understanding the use of MICT in other business contexts. Future research is expected to focus more deeply on the relationships between individual factors and on assessing their relative influence on user behaviour. Research into clinical settings outside hospitals is another area of potential future work, especially as some of these, such as use in emergency ve-

hicles, involve distinctive forms of mobility that may provide additional insights into the usage of MICTs. It is also planned to extend the framework to non-healthcare settings.

ACKNOWLEDGMENTS

Funding for this research was received from Fundação para a Ciência e Tecnologia, Lisbon (BD/8121/2002), and St. Edmunds College, Cambridge. The support of all those involved in obtaining access to the study sites as well as all staff at those sites is gratefully acknowledged.

REFERENCES

- Barret, J. R., Strayer, S. M., & Schubart, J. R. (2004). Assessing medical residents' usage and perceived needs for personal digital assistants. *International Journal of Medical Informatics*, 73(1), 25-34.
- Brenda, M., & Gadd, S. C. (2001). Introducing handheld computing into a residency program: Preliminary results from qualitative and quantitative inquiry. *Proceedings of the AMIA Symposium* (pp. 428-432).
- Cox, J. (2002). *Networked mobile devices help improve patient care and diagnosis*. Retrieved from <http://www.nwfusion.com/research/2002/1209sector.html>
- Davis, G. B. (2002). Anytime/anyplace computing and the future of knowledge work. *Communications of the ACM*, 45(12), 67-73.
- Dryer, D. C., Eisbach, C., & Ark, W. S. (1999). At what cost pervasive? A social computing view of mobile computing systems. *IBM Systems Journal*, 38(4), 652-675.
- Kelly, J. (2001). Going wireless. *Hospital Health Networks*, 74(11), 65-66, 68.
- Kristoffersen, S., & Ljungberg, F. (2000). *Mobility: From stationary to mobile work. Planet Internet*. Lund: Studentlitteratur.
- Luff, P., & Heath, C. (1998). Mobility in collaboration. *Proceedings of the ACM 1998 Conference on Computer Supported Cooperative Work*, Seattle, WA.
- Martins, H. M. G., & Jones, M. R. (2005). What's so different about mobile information communication technologies (MICT) for clinical work practices: A review of selected pilot studies. *Health Informatics Journal*, 11, 123-134.
- Martins, H. M. G., Nightingale, P., & Jones, M. R. (2005). Temporal and spatial organisation of doctors' computer usage in a UK hospital department. *Medical Informatics & the Internet in Medicine*, 8(2), 135-142.
- McLeod, T. G., Ebbert, J. O., & Lymp, J. F. (2003). Survey assessment of personal digital assistant use among trainees and attending physicians. *Journal of the American Medical Association*, 10(6), 605-607.
- Riley, R. T., Lorenzi, N. M., & Dewan, N. A. (2002). Barriers and resistance to informatics in behavioral health care. In N. A. Dewan, R. T. Lorenzi, & S. R. Bhattacharya (Eds.), *Behavioral healthcare informatics* (pp. 140-148). New York, Springer-Verlag.
- Westbrook, J. I., Gosling, A. S., & Coiera, E. (2004). Do clinicians use online evidence to support patient care? A study of 55,000 clinicians. *Journal of the American Medical Informatics Association*, 11(2), 113-120.

ENDNOTES

- ¹ We define clinical work practices as doctors' work directly related to patient care (e.g., prescribing a drug or viewing an X-

ray for treatment plan decision making). Although equally interesting, doctors' activities such as teaching, research, or all those occurring "outside" hospital working hours are not considered in this chapter.

² "Jobs" corresponds to discreet tasks to be carried out only by doctors, for example,

drawing blood, changing an infusion, doing a small medical procedure. Due to the intensive care aspect of this department, these were numerous and new ones could be required throughout the day and night as the conditions of the patients evolved.

Chapter VI

Mobility in Healthcare for Remote Intensive Care Unit Clinical Management

Carolyn McGregor

University of Western Sydney, Australia

ABSTRACT

This chapter reviews current research directions in healthcare mobility and assesses its impact on the provision of remote intensive care unit (ICU) clinical management. Intensive care units boast a range of state of the art medical monitoring devices to monitor a patient's physiological parameters. They also have devices such as ventilators to offer mechanical life support. Computing and IT support within ICUs has focused on monitoring the patients and delivering corresponding alarms to care providers. However many intensive care unit admissions are via intra and inter health care facility transfer, requiring receiving care providers to have access to patient information prior to the patient's arrival. This indicates that opportunities exist for mobile gadgets, such as personal digital assistants (PDAs) to substantially increase the efficiency and effectiveness of processes surrounding healthcare in the ICUs. The challenge is to transcend the use of these mobile devices beyond the current usage for personal information management and static medical applications; also to overcome the challenges of screen size and memory limitations. Finally, the deployment of mobile-enabled solutions within the healthcare domain is hindered by privacy, cost and security considerations and a lack of standards. These are some of the significant topics discussed in this chapter.

INTRODUCTION

Intensive care units (ICUs) worldwide offer support for patients in need of critical care. The research, development, and adoption of new information technologies (ITs) and information

systems (ISs) within ICUs, and particularly neonatal intensive care units (NICUs) to support patient and care provider mobility, is currently lagging behind other industries and other areas of healthcare (McGregor, Heath, & Wei, 2005a; McGregor, Kneale, & Tracy, 2005b;

Wu, Wang, & Lin, 2005). In order to understand and improve upon this lag, we need to understand the current healthcare scenario within the context of intensive care.

To start with, mobile clinical management solutions within the context of intensive care units need to consider not only the mobility of the patient, but equally importantly, the mobility of the care provider. When an incident requiring critical care occurs, patients may already be located in the care provider's ICU. However, the patient may also be located elsewhere in the care provider's hospital, in another hospital, in their home, or in another location outside the hospital of the care provider. Patients may also be in transit between any of these locations via an ambulance, helicopter, or inter-hospital transport.

Care providers can be understood in this chapter as any physician, clinician, or nursing specialist responsible for some aspect of the clinical management of the ICU patient. In daily routines, physicians, clinicians, nurses, and other staff of the hospital have to be reached and updated of new incidents and information while they are commuting in their work environments (Kafeza, Chiu, Cheung, & Kafeza, 2004). However, similar to the patients, care providers may also be located within the ICU, their office, elsewhere within their hospital, in their home, or in another location outside their hospital (e.g., attending an off-site meeting or conference).

Ammenwerth, Buchauer, Bludau, and Haux (2000) report that one of the major clinical management issues that mobile technologies can help with within the hospital is communication and reachability of care providers. This clinical management issue has the additional challenge of determining to whom the message should be sent (Kafeza et al., 2004). Both of these issues are particularly relevant within the ICU setting.

When critical care clinical management is required, the sooner the patient/care provider(s) information exchange can commence, the faster the clinical management can commence. In the case where the patient and care provider(s) are not located together within the ICU, critical care can still commence, provided there is adequate clinical management support to facilitate clinical decision making and execution.

Mobile healthcare systems (MHSs) have been defined by Wu et al. (2005) as the use of IS/IT to exchange healthcare information and services via mobile devices anytime and anywhere, providing patients and care providers with easy access to resources whether stationary or moving.

Recent research directions for computing and IT support within ICUs has focused on the delivery of alarms/alerts to care providers (Catley & Frize, 2003; Catley, Frize, Walker, & St. Germain, 2003; Shabot, LoBue, & Chen, 2000; Sukuvaara, Makivirta, Kari, & Koski, 1989; van der Kouwe & Burgess, 2003). However these approaches do not enable mobility in healthcare and neither do they exploit the substantial benefits possible by proper application of mobility. Furthermore, many intensive care unit admissions are via intra- and inter-healthcare facility transfer, requiring receiving care providers to have access to patient information, prior to the patient's arrival and often while the care provider is also in transit. These are some interesting challenges in terms of communication and reachability of care providers.

Recent surveys show that between 25-35% of physicians, as distinct from care providers in general, use personal digital assistants (PDAs) (Carroll & Christakis, 2004; Fontelo, Kim, & Locatis, 2003). However, Carroll et al. (2003) further note that these PDAs are mainly for personal information management and static medical applications. Opportunities exist for PDAs and similar handheld devices to enhance

and effectively deliver services within the ICU clinical management sector. However, PDA screen size and memory are seen as crucial factors in the development of PDA applications. In addition, deployment of mobile-enabled solutions within the healthcare domain is impacted by privacy, confidentiality, cost, and security considerations in addition to a current lack of standards.

This chapter reviews current research directions in healthcare mobility and assesses its impact on the provision of remote intensive care unit clinical management. A background to intensive care unit clinical management is first introduced. Recent computing and IT-related research to support ICUs is then presented. Hardware and associated research to support ICU clinical management mobility is then described. A comparison of recent ICU research within the context of its ability to support mobility is presented. Issues impacting the implementation of mobile ICU clinical management solutions are then detailed. Finally, the conclusion and future directions are presented.

BACKGROUND TO ICU CLINICAL MANAGEMENT

Clinical management systems are designed to assist care providers in diagnosis and treatment using existing, already established methods of diagnosis and accepted treatments (Gross-Portney & Watkins, 2000). Hence, mobility in clinical management must support mobility in relation to diagnosis and treatment. Tasks such as medication monitoring, emergency hospitalization of patients, laboratory examination results, ordering and shipment of drugs, and exchange of information relating the patient clinical management occur frequently (Kafeza et al., 2004). Within the ICU context, clinical management systems must respond actively

and very timely to the patient's needs, which can be life critical.

One of the most prominent objectives within the modern hospital is the need for accurate, safe, and continuous communications among departments and highly specialized medical staff. In addition, the need for flexible communications to enable communication with other hospitals is also dominant. As a result there has been a great demand among the care providers for a mobile alert management system that is robust, efficient, cost effective, simple, and user friendly (Kafeza et al., 2004).

Intensive care units (ICUs) worldwide offer support for patients in need of critical care. They boast a range of state-of-the-art medical monitoring devices to monitor a patient's physiological parameters such as blood oxygen, blood pressure, and heart rate. Other devices such as ventilators offer mechanical life support.

Broadly, there are three types of intensive care, namely, adult, pediatric, and neonatal. While the age of the patient is the differentiator between adult and pediatric ICUs, the clinical management differs greatly from these ICUs to the neonatal ICUs (NICUs)—where gestational age greatly impacts clinical management.

Approximately 18% of babies born in New South Wales (NSW), Australia, require special care or neonatal intensive care admission (NSW Health Department, 1994). Premature babies can be up to 17 weeks early and may only weigh 450 grams; they can spend three or four months in intensive care and have dozens of specific diseases before discharge. In addition, 15% of neonatal intensive care admissions are transferred after delivery from smaller remote hospitals without intensive care facilities. Similar conditions apply elsewhere within Australia and internationally, where small remote hospitals are spread throughout a given country supported by centrally located referral hospitals with NICUs.

Remote hospitals have equipment to provide limited NICU support within ‘special care nurseries’; but without the ability for a neonatologist to receive information from this equipment, the baby must be moved to a referral hospital with neonatologist support. Given the critical requirement to maintain a consistent environment, moving a baby at this time can be life threatening. Critically ill, term and pre-term babies that have to be transferred have higher mortality rates and much higher rates of long-term disability than similar babies born in hospitals with intensive care facilities (McGregor, Bryan, Curry, & Tracy, 2002). A major limitation is that the attending care provider at the remote hospital must contact a neonatal specialist (neonatologist) via telephone, or in some instances the provider—who may or may not be located at the NICU at that time—must describe via e-mail (Deodhar, 2002) the baby’s symptoms and, where possible, relay any physiological information verbally, or narratively in the context of an e-mail. The consulting neonatologist must then make decisions based on this verbal or textual exchange.

It is very common for critically ill babies to have significantly abnormal variation in the measured parameters minute by minute, and not all these variations are made available to the consulting neonatologist. Frequent transient falls in blood pressure and blood oxygen content, often with swings into the high range, may be of critical importance in survival and quality of survival, free of significant disability (Lister, Bryan, & Tracy, 2000).

Hence the neonatologists located at referral hospitals require the ability to obtain information from the monitors attached to babies. Similarly, a neonatologist need not be located at a PC within the hospital to view patient data, but should be free to view this information through any device offering a secure Internet/intranet

connection. These scenarios open up opportunities for application of mobility in ICU management, as discussed in this chapter.

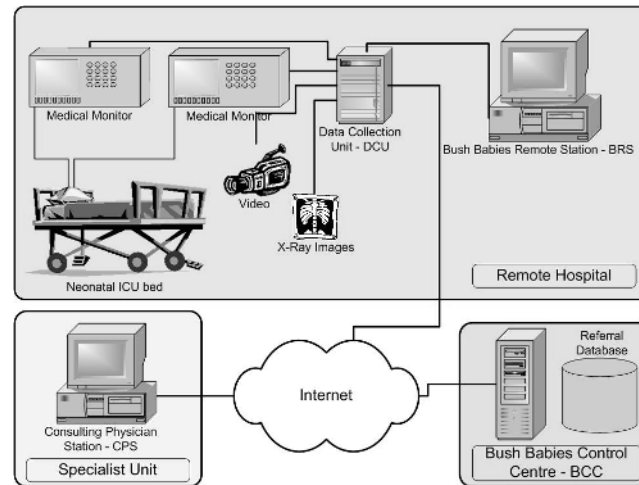
RECENT COMPUTING AND IT RESEARCH TO SUPPORT ICU CLINICAL MANAGEMENT

Much of the recent computing and IT related research to support intensive care units (ICUs) has focused on clinical alerts (Catley & Frize, 2003; Catley et al., 2003; Shabot et al., 2000; Sukuvaara et al., 1989; van der Kouwe & Burgess, 2003). The information made available to these systems is limited to a small set of physiological data and/or clinical data from patients located within their ICUs. In addition, care provider access to these systems is limited to the receipt of alerts, with minimal content via e-mail and in some cases pagers.

An integrated XML-based healthcare framework for NICU clinical alerts is described by Catley and Frize (2003) and Catley et al. (2003). The alerting is based on the individual device alarms, and predicts mortality and ventilation requirement probability and estimated length of stay. In the implementation as presented, only e-mail alerts and Java GUI alerts are generated. They indicate that future directions for this research are a WML-based alert using Java Servlet technology running on a Web server connected to a WAP Gateway. The WAP Gateway would transmit alerts via WML to mobile devices.

Shabot et al. (2000) describe a software system which extracts clinical information from clinical information systems on a continuous basis and sends it through event detection algorithms. Alerts for detected events are forwarded through a commercial paging system to designated care providers and pharmacists.

Figure 1. Bush Babies architecture (McGregor et al., 2004)



InCare was a rule-based alarming system prototype to support clinical management within ICUs (Sukuvaara et al., 1989) that detected four important patient pathological conditions, which develop gradually during postoperative recovery of cardiac patients, namely: (1) hyperdynamic state, (2) hypoventilation, (3) hypovolemia and left ventricular failure. User interface to the alarms was via a PC.

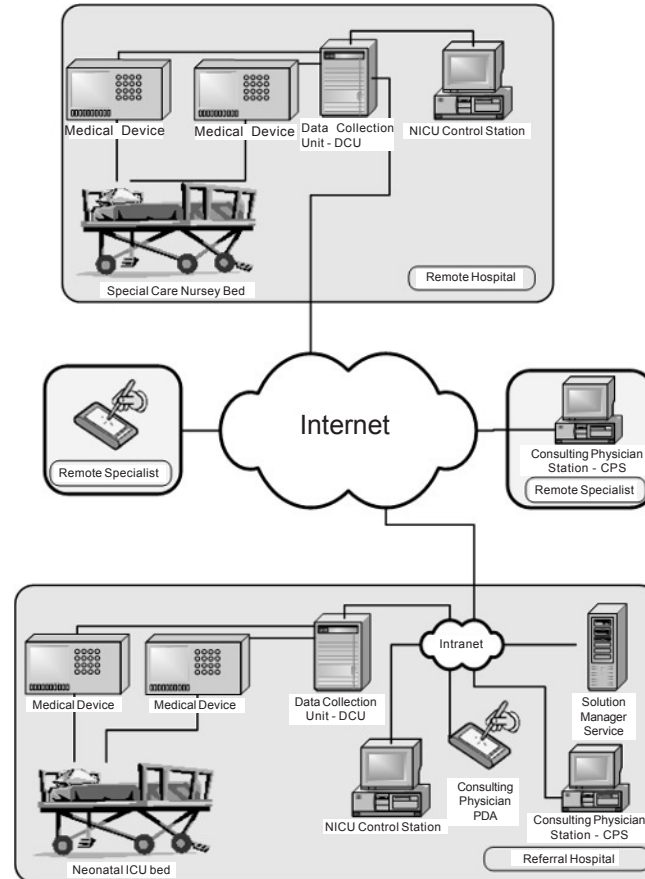
Van der Kouwe and Burgess (2003) present an architecture for continuous electrophysiological monitoring within a neurointensive care unit. While this information is available in real time through Web-based interface, access is limited to PCs via the hospital's secure intranet.

Shin, Huh, Lee, and Kim (2003) have developed Web-based real time for checking temperature and humidity within infant incubators in NICUs. However the access to this information was limited to the NICU central monitoring station PC, with the information delivered via the hospital intranet.

The Bush Babies Broadband research project (McGregor, Kneale, & Tracy, 2005b), supported by the Telstra Broadband Fund, aims to

significantly improve the quality of treatment for babies born in regional and remote areas by providing the first on-demand virtual neonatal intensive care unit architecture in Australia. The Bush Babies architecture is shown in Figure 1. Real-time data collected from medical monitors and ventilators attached to the baby, audiovisual streams, and static physiological data such as x-ray images are transmitted from the data collection unit (DCU) to the consulting neonatologist to gain a better picture of the patient's condition than is currently available. The Bush Babies Remote Station enables the remote care provider to initiate a bush babies session. The BRS uses a centralised Bush Babies Control Centre database to select a neonatologist from a NICU where space for the baby would be available if transport was required. A limitation of that research is that the only consulting physician station (CPS) device that neonatologists can use to view this patient condition information is via the screen of a PC or laptop. Given that consulting neonatologists are not always located in their offices within their NICU when there services are required,

Figure 2. e-Baby architecture (McGregor et al., 2005a)



alternate and more portable technologies to deliver the information need to be investigated.

In association with the Bush Babies project, McGregor et al. (2005a) and McGregor, Purdy, and Kneale (2005c) propose a Web service-based framework for the transmission of XML-encoded physiological data output by medical monitoring and life support devices. That research, together with the previously mentioned Bush Babies, are portions of the "e-Baby" research collaboration (McGregor et al., 2002, 2005a, 2005b, 2005c) that is researching new approaches to the application of computing and information technology to support mobility in healthcare through local and remote neonatal intensive care. The high level e-Baby architecture is shown in Figure 2.

The Web service-based framework for the transmission of XML-encoded physiological data forms part of the solution manager service (SMS). The SMS is situated in the referral hospital NICU and receives and stores data collected by the data collection units via the *physiological log Web service* for near-real-time analysis and trend detection. The consulting physician station is used by the neonatologists to access the physiological data located in SMS via a set of *analyse Web services* or via a direct link to the data as it is streamed through the *physiological log Web service*. That research enables patient and care provider mobility, as a result of the Web service-based data transmission. However, the prototype as presented only tests transmission of data from

patients located either within the NICUs or remote hospital and care providers located within the hospital with the NICU. In addition, care provider access is only available through the PC-based CPS, though future research indicates the forwarding of data to consulting physician PDAs.

HARDWARE TO SUPPORT ICU CLINICAL MANAGEMENT MOBILITY

The variety of mobile devices available for use to support clinical management within ICUs include: PDAs, laptops, notebooks, GPSs, and smartphones. Recent surveys show that approximately 25% or more of physicians use PDAs, although mainly for personal information management and static medical applications. A PDA provides many advantages. It starts quickly at the push of a button. It is convenient to carry around, fitting easily into a shirt pocket or handbag. Some devices can function for weeks of regular use after a quick battery recharge. However, PDA screen size and memory are seen as crucial factors in the development of PDA applications (Fontelo et al., 2003).

Carroll and Christakis (2004) recently surveyed pediatricians in relation to their use of PDAs and found that 35% currently use PDAs at work; the most common uses were for drug reference (80%), personal scheduling (67%), and medical calculations (61%).

A plethora of software packages are available for PDAs to support clinical management in the broad areas of medical handheld software collections, medical publishers, university medical handheld resources, document readers, access to medical literature, pharmacopoeias, specialty specific, and patient tracking. Patient

tracker (www.handheldmed.com) gives the user the option to enter patient records, including demographics, laboratory results, medication, allergy lists, test results, and radiology reports. Wardwatch (www.torlesse.com) was designed to aid medical staff in ward rounds. Medical Pocket Chart (www.gemedicalsystems.com) is an electronic medical record keeper (Fischer, Mehta, Wax, & Lapinsky, 2003).

Providing care providers with the ability to perform research at the bedside via PDA access to PubMed and clinical trial Internet sites was described by Fontelo et al. (2003).

Carroll (2001, 2002) and Carroll, Tarczy-Hornoch, O'Reilly, and Christakis (2004) describe the implementation of a PDA-based patient record and charting system for an NICU; however, the charting component relies on the care providers completing patient flowsheets manually into the PDA, rather than having the physiological data streaming to the PDA.

A PDA-based approach for managing patient data is defined by Lapinsky et al. (2001). Patient data was entered into the Memopad using a customized template. This data was transferred between care providers using the PDAs' infrared ability. Daily paper notes were generated via infrared link to a HP LaserJet printer, as the hospital policy required paper records.

NICU Notes (Schulman, 2003) enables care providers to collect data at the point of care and utilizes synchronization to move the patient data from the PDA via an ODBC DSN to a secure Microsoft Access application.

A critical PDA issue as defined by Carroll (2002) is the asynchronous nature of hot syncing. Information on the PC and PDA only match immediately after hot syncing. The sheer volume of data being passed during the hot sync process caused the hot sync process to fail

Table 1. Broad functionality

Research Identifier	Clinical Management Function(s) Supported	Decision Time	Decision Quality
(Catley & Frize, 2003; Catley et al., 2003)	Clinical alerts predicting mortality, ventilation requirements, and length of stay	Non-critical	Not available
(Shabot et al., 2000)	Clinical event detection, laboratory results, and medication alerts	Non-critical and life threatening	High
(Sukuvaara et al., 1989)	Alarms relating to postoperative recovery of cardiac patients	Life threatening	High
(van der Kouwe & Burgess, 2003)	Samples electro physiological data	Life threatening	High
(Shin et al., 2003)	Infant humidicrib temperature and humidity monitoring	Life threatening	High
(McGregor et al., 2005b)	Video, image, and physiological data stream monitoring	Life threatening	High
(McGregor et al., 2005a, 2005c)	Physiological data stream monitoring	Life threatening	High
(Carroll, 2002; Carroll et al., 2004)	Patient record charting system	Non-critical	Medium
(Lapinsky et al., 2001)	Managing patient data	Non-critical	Medium
(Schulman, 2003)	Managing patient data	Non-critical	Medium

intermittently, resulting in incomplete or duplicated information.

COMPARISON OF RECENT RESEARCH

Having discussed the routine usage of mobility in healthcare thus far, we now focus our attention on comparing the previously presented computing and IT research to support ICU clinical management—with an aim to assess the ability of that research to support mobility.

The comparison is broadly categorized into the areas of broad functionality, patient and care provider mobility, and finally architecture mobility. Within the context of broad function-

ality, three areas were considered: the clinical management function(s) that was (were) being supported, the time sensitivity of the decision, and the decision quality sensitivity. Time sensitivity was considered important, as Panniers (1999) has stated that only low and medium urgency decisions are suitable for computerization into a decision support system to support clinical management. The comparison based on broad functionality is presented in Table 1.

Secondly, the degree of patient and care provider mobility was assessed, together with the extent of information that was available to the care provider and is summarized in Table 2.

Finally, the extent to which the proposed architectures incorporate mobility is summarized in Table 3.

Mobility in Healthcare for Remote Intensive Care Unit Clinical Management

Table 2. Degree of patient and care provider mobility

Research Identifier	Extent of Patient Mobility	Extent of Care Provider Mobility	Extent of Information Available to (from) Care Provider
(Catley & Frize, 2003; Catley et al., 2003)	Located within NICU	Accessing PC within hospital (wireless access proposed for future research)	Text-based notification that physiological data values have exceeded threshold
(Shabot et al., 2000)	Located within ICU	Alerts sent to pager	Limited text-based alerts, based on type of alert
(Sukuvaara et al., 1989)	Located within ICU	Accessing PC within ICU	Physiological data streams
(van der Kouwe & Burgess, 2003)	Located within ICU	Accessing PC within hospital	Physiological data streams
(Shin et al., 2003)	Located within NICU	PC located at NICU control station	Data stream
(McGregor et al., 2005b)	Located within local NICU or remote hospital	Accessing PC within hospital	Video, image, and physiological data stream
(McGregor et al., 2005a, 2005c)	Located within local NICU or remote hospital	Accessing PC within hospital	Physiological data stream
(Carroll, 2002; Carroll et al., 2004)	Located within NICU	Located within NICU	(Clinical charts updated by care provider)
(Lapinsky et al., 2001)	Located within ICU	Located within ICU	(Clinical charts updated by care provider)
(Schulman, 2003)	Located within NICU	Located within NICU	(Clinical charts updated by care provider)

Table 3. Extent to which architectures incorporate mobility

Research Identifier	User Interface Device(s)	Networks	Software/Middleware
(Catley & Frize, 2003; Catley et al., 2003)	Web-based	Wired hospital intranet	XML-based messaging
(Shabot et al., 2000)	Pager	PageNet network	Pager messaging
(Sukuvaara et al., 1989)	PC GUI	Wired hospital intranet	Not available
(van der Kouwe & Burgess, 2003)	Web-based	Wired hospital secure intranet	Not available
(Shin et al., 2003)	Web-based	Wired hospital secure intranet	HTML document
(McGregor et al., 2005b)	Web based	Wired hospital secure intranet	Simple Medical Data Protocol (SMDP) document
(McGregor et al., 2005a, 2005c)	Web-based	Wired hospital secure intranet	XML document
(Carroll, 2002; Carroll et al., 2004)	PDA	Hotsync to PC	PDA HotSync
(Lapinsky et al., 2001)	PDA	Between PDAs and printer via infrared	PDA infrared
(Schulman, 2003)	PDA	Hotsync	PDA Hotsync

Contrary to Panniers' (1999) time sensitivity observations, the research indicates that clinical management systems are being developed to support life-threatening conditions by alerting care providers quickly of the development of the situation.

When data is being forwarded directly in a time series stream from medical devices, data quality is high. However, when PDAs are used to collect data from care providers, errors still occur.

None of the research reviewed catered to patients located elsewhere within the care provider's hospital, and the only out-of-hospital location that was supported was another special care nursery within McGregor et al.'s (2005b) research.

Only Shabot et al. (2000) enabled care provider communication from outside the hospital, and this was via a pager. Hence, information available to care providers outside the hospital was limited.

While most user interfaces for the delivery of information were Web enabled, the task of delivery to devices other than PCs and notebooks has not been adequately addressed.

Standards for user interfaces and communications have not been developed as part of these research efforts, nor does the research contain references to other standards efforts. This indicates that such standards do not exist within the context of ICU clinical management.

ISSUES IMPACTING IMPLEMENTATION

In addition to considering PDA usage, there are several factors that to date are still issues impacting successful research, development, and implementation of mobile clinical management solutions within the ICU setting. These factors are indeed common to all clinical man-

agement solutions and include wireless network interference in ICUs, security, privacy, confidentiality, cost, and a lack of communications standards.

While the adoption of 802.11b wireless networks is increasing, Fontelo et al. (2003) state that security issues may preclude current deployment of wireless devices for medical data access and utilize infrared access stations within their research. This link was restricted to a 15-degree arc on each side of the centre and a maximum distance of 8 feet. They also found that a transient, rapid disruption of the infrared (IR) beam, such as a person walking between the IR point and the PDA, did not disconnect an established link. In addition, the use of devices via wireless networks within ICUs currently interferes with many of the devices used within ICUs for critical care.

Fischer et al. (2003) state that patient confidentiality and costs are the main implementation issues for handheld systems used within clinical management and that consumer interest may be the limiting factor to successful implementation. In addition, the lack of standards and limited bandwidth for data transfer may also impact increased implementation.

However, they conclude that *a growing body of literature supports the use of handheld devices in a variety of medical settings and with the rapid advances in this technology, the mobile computer may well become an essential medical tool.*

While the deployment of mobile clinical management solutions within ICUs offers the potential of improved patient care and service quality and increasing care efficiency, most applications within the broad healthcare context have failed or not been implemented as predicted, with 30% of failures attributed to non-technical factors (Wu et al., 2005).

For mobile IS/IT solutions to significantly impact ICU clinical management and result in a

paradigm shift in the approaches to healthcare within this context, several standards need to be developed to support the ICU clinical management functions. These functions include the exchange of information relating to the patient clinical management, medication monitoring, ordering and shipment of drugs, and the examination of laboratory results.

While all monitoring and life support devices used within ICUs to supply physiological data have the ability to output the device readings usually via a serial port, the data formats vary greatly from device to device. As such, efforts to make device data from devices attached to local or remote patients accessible for mobile viewing are hampered by the myriad of formats required for transmission. McGregor et al. (2005a, 2005c) propose a Web service-based framework for the transmission of physiological data output by such devices, proposing an XML format for such data transmission; however, it has not to date resulted in mainstream standards adoption for the transmission of physiological data.

To enable ICU clinical management mobility, standards for physiological data transmission, medication monitoring, laboratory examination results, and the ordering and shipment of drugs are still required.

CONCLUSION AND FUTURE DIRECTION

While the paradigm shift to mobile clinical management for ICUs offers the potential to significantly improve the speed, efficiency, and effectiveness of critical care within ICUs, there are currently several factors impacting its mainstream adoption.

This chapter has presented a review of current healthcare mobility within the context

of ICU clinical management. A comparison of recent ICU computing and IT-related research indicates that the issue of both patient and care provider mobility has not been considered a priority within these research efforts.

Several issues continue to impact successful implementation of mobile clinical management solutions for ICUs; these include issues relating to the use of wireless networks, in addition to security, privacy, cost, and a lack of communications for data exchange and user interface standards.

Current computing and IT-related research to design and develop the next-generation IT/IS solutions for ICUs is not adequately incorporating the issue of mobility of patient and care provider.

Traditionally, clinical research is used as the catalyst for providing evidenced-based recommendations for change to clinical management practices. To gain care provider acceptance, changes to clinical management practices proposed through the introduction of computing and IT approaches should consider traditional clinical research approaches to validate findings and gain support and acceptance.

ACKNOWLEDGMENTS

The work presented in this chapter is associated with a larger, long-term research effort known as “e-Baby” (McGregor et al., 2002, 2005a, 2005b, 2005c) that is researching new approaches to the application of computing and information technology to support mobility in healthcare through local and remote neonatal intensive care. The researchers wish to thank the staff of the Neonatal Intensive Care Unit, Nepean Hospital, Sydney West Area Health Services, NSW, for their continued support and collaboration in the e-Baby research.

REFERENCES

- Ammenworth, E., Buchauer, A., Bludau, B., & Haux, R. (2000). Mobile information and communication tools in the hospital. *International Journal of Medical Informatics*, 57(1), 21-40.
- Carroll, A. E., Tarczy-Hornoch, P., O'Reilly, E., & Christakis, D. A. (2004). The effect of point-of-care personal digital assistant use on resident documentation discrepancies. *Pediatrics*, 113(3), 450-454.
- Carroll, A. E., & Christakis, D. A. (2004). Pediatricians' use of and attitudes about personal digital assistants. *Pediatrics*, 113(2), 238-242.
- Carroll, S. S., & Tarczy-Hornoch, P. (2001, November 6). Development of a personal digital assistant (PDA) based client/server NICU patient data and charting system. *Proceedings of the AMIA Symposium*, Washington, DC (pp. 100-104).
- Carroll, S. S., & Tarczy-Hornoch, P. (2002, November 12). The implementation of a personal digital assistant (PDA) based patient record and charting system: Lessons learned. *Proceedings of the AMIA Symposium*, San Antonio, TX (pp. 111-115).
- Catley, C., & Frize, M. (2003, April 9-12). A prototype XML-based implementation of an integrated 'intelligent' neonatal intensive care unit. *Proceedings of the 4th International IEEE EMBS Special Topic Conference on Information Technology Applications in Biomedicine*, Birmingham, UK (pp. 322-325).
- Catley, C., Frize, M., Walker, C. R., & St. Germain, L. (2003, September 17-21). Integrating clinical alerts into an XML-based health care framework for the neonatal intensive care unit. *Proceedings of the 25th Annual IEEE International Conference of the Engineering in Medicine and Biology Society*, Cancun, Mexico (Vol. 2, pp. 1276-1279).
- Deodhar, J. (2002). Telemedicine by e-mail: Experience in neonatal care at a primary care facility in rural India. *Journal of Telemedicine and Telecare*, 8(2), 20-21.
- Fischer, S. T., Mehta, S., Wax, R., & Lapinsky, S. E. (2003). Handheld computing in medicine. *JAMIA*, 10(2), 139-149.
- Fontelo, A. M., Kim, G., & Locatis, C. (2003). The PDA as a portal to knowledge sources in a wireless setting. *Telemedicine Journal and E-Health: The Official Journal of the American Telemedicine Association*, 9(2), 141-147.
- Gross-Portney, L., & Watkins, M. (2000). *Foundations of clinical research, applications to practice* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.
- Kafeza, E., Chiu, D. K. W., Cheung, S. C., & Kafeza, M. (2004). Alerts in mobile healthcare applications: Requirements and pilot study. *IEEE Transactions on Information Technology in Biomedicine*, 8(2), 173-181.
- Lapinsky, S. E., Weshler, J., Mehta, S., Varkul, M., Hallett, D., & Stewart, T. E. (2001). Handheld computers in critical care. *Critical Care*, 5(4), 227-231.
- Lister, R., Bryan, G., & Tracy, M. (2000, July 3-5). The e-Babies Project: Integrated data monitoring and decision making in neonatal intensive care. *Proceedings of the European Conference of Information Systems*, Vienna, Austria (Vol. 2, pp. 1289-1293).
- McGregor, C., Bryan, G., Curry, J., & Tracy, M. (2002, January 7-10). The e-Baby data warehouse: A case study. *Proceedings of the*

35th Annual Hawaii International Conference on System Sciences (HICSS), Big Island (pp. 3018-3024).

McGregor, C., Heath, J., & Wei, M. (2005a, March 29-April 1). A Web service based framework for the transmission of physiological data for local and remote neonatal intensive care. *Proceedings of the IEEE International Conference on E-Technology, E-Commerce and E-Service*, Hong Kong (pp. 496-501).

McGregor, C., Kneale, B., & Tracy, M. (2005b, July 4-7). Bush Babies broadband: On-demand virtual neonatal intensive care unit support for regional Australia. *Proceedings of the 3rd International Conference on Information Technology and Applications (ICITA)*, Sydney, Australia (Vol. 2, pp. 113-117).

McGregor, C., Purdy, M., & Kneale, B. (2005c). Compression of XML physiological data streams to support neonatal intensive care unit Web services. *Proceedings of the IEEE International Conference on E-Technology, E-Commerce and E-Service*, Hong Kong (pp. 486-489).

NSW Health Department. (1994). *NSW midwives data collection*.

Panniers, T. L. (1999, November 6-10). Selecting appropriate problems for decision support system in the neonatal intensive care unit. *Proceedings of the AMIA Annual Symposium*, Washington, DC (p. 1133).

Schulman, J. (2003, November 8-12). NICU Notes: A Palm OS(R) and Windows(R) database software product and process to facilitate patient care in the newborn intensive care unit. *Proceedings of the AMIA Symposium*, Washington, DC (p. 999).

Shabot, M. M., LoBue, M., & Chen, J. (2000, January 4-7). Wireless clinical alerts for critical medication, laboratory and physiologic data. *Proceedings of the 33rd Annual Hawaii International Conference on System Sciences*, Maui, HI (Vol. 1, p. 6).

Shin, D. I., Huh, S. J., Lee, T. S., & Kim, I. Y. (2003). Web-based remote monitoring of infant incubators in the ICU. *International Journal of Medical Informatics*, 71, 151-156.

Sukuvaara, T., Makivirta, A., Kari, A., & Koski, E. (1989, September 19-22). An intelligent intensive care alarming system. *Proceedings of the Conference on Computers in Cardiology* (pp. 225-228).

van der Kouwe, A. J. W., & Burgess, R. C. (2003). Neurointensive care unit system for continuous electrophysiological monitoring with remote Web-based review. *IEEE Transactions on Information Technology in Biomedicine*, 7(2), 130-140.

Wu, J.-H., Wang, S.-C., & Lin, L.-M. (2005). What drives mobile health care? An empirical evaluation of technology acceptance. *Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS '05)*.

Chapter VII

The Evaluation of Wireless Devices Used by Staff at Westmead Hospital, Sydney

Sandra Synthia Lazarus
University of Sydney, Australia

ABSTRACT

This chapter reports on a study to research and evaluate the use of latest generation wireless devices—typically personal digital assistant devices (PDAs)—by clinical staff at the large Westmead Hospital located in the west of Sydney, Australia. Currently, medical reports in this and other hospitals are primarily recorded on paper supported by personal computers at nursing stations. However, there is very little or no access to medical reports and decision-making tools for medical diagnosis at the patient's bedside—the precise location at which most medical decision-making occurs. Delays in access to essential medical information can result in an increased time taken for accurate diagnosis and commencement of appropriate medical management of patients. This chapter discusses the application of hand held devices into more powerful processing tools connected to a centralised hospital data repository that can support medical applications.

INTRODUCTION

Currently, personal digital assistant (PDA) devices are used by most junior hospital staff to view formulary information. A short survey carried out showed that almost 90% already own a PDA device and use it for drug and medication referencing. Access to drug infor-

mation databases on PDAs provides a fast and easy way to reference medication details including interactions. Hospital staff members also use PDA devices for day-to-day calendar and appointment notification (Carroll & Cristakis, 2004). But it is just recently, due to the evolution of PDA devices, that it has become possible to take the next step and utilise

the full capacity of the PDA devices in a hospital environment.

The study was conducted with a variety of staff members in the hospital; these included senior clinical staff, specialist staff, and junior staff. This was from across the hospital—from general wards, special units, and the emergency department. The duration of this trial was for nine months, with a rotation of staff members. Staff members were assessed before becoming part of the trial as to how long it takes the individual to carry out a particular task without the assistance of a PDA device; these tasks were timed. Then a complete instruction booklet was given to the hospital staff with the PDA device to make use of as they found it necessary in their daily duties.

After a certain period when the staff members were comfortable in using the device, time was again measured for different tasks with and without PDA devices. This set the background for the trial. To evaluate the social accepts of the trial, a survey was conducted to gather the thoughts of the users on the functionality and the useability of the device in a hospital environment.

THE EVOLUTION OF PDA DEVICES

Over the last few years, the PDA has evolved in ways originally unimaginable. One of the first PDAs to enter the market was the Palm Pilot (now more commonly known as the Palm). When first released, the Palm came with a monochrome screen and had a limited amount of onboard memory. Since then the Palm evolved into a PDA that featured expandable slots for additional memory and expandability options (i.e., modems, printers, and other peripherals). In addition, Palm PDAs now feature high-resolution colour screens which are continu-

ously improving. Furthermore, the Palm also comes with built-in wireless access that permits it to connect to various types of devices and networks, allowing it to be truly portable (as originally intended).

Pocket PCs entered the market shortly after Palm PDAs were introduced; however, unlike Palms the Pocket PC is more of a PDA standardisation rather than a brand name. Pocket PCs are made by several companies, for example Dell, Toshiba, and HP/Compaq; however, the core operating system is written by Microsoft. On the other hand, the Palm PDA operating system (Palm OS) is maintained by Palm itself, which ensures that all features of the operating systems are fully functional. Small distinctions such as this have allowed both devices/brands to compete for the same market space and are still considered independent niche markets.

Since Palm has been on the market for a considerably longer period of time, there is a remarkable amount of software support found throughout the Internet. Palm PDAs now have the ability to operate seamlessly with Windows, Linux, UNIX, Novell, Macintosh, and practically any type of network. There is a considerable amount of open source development done within the communities that make featured software available for Palms. Similarly, Pocket PCs are supported through a large number of Internet sites. Since standards are maintained, which include desktop and server operating systems, these have the ability to ensure that their operating systems designed for Pocket PCs will meet the same quality control and compatibility assurance. Nevertheless, with the rivalry between these two PDA vendors, it is possible to see further improvements such as superior screen resolution, improved standard of expandability options, inclusions of newer wireless technology such as 802.11g, and most importantly reduced costs, allowing more people

Figure 1. The Palm Tungsten T3 has the following features: Bluetooth technology synchronisation, 64 MB memory, stretch display, portrait and landscape, moderate resolution 320×480 pixels (Palm One, 2005)



to access such technology. Since their evolution, PDA devices can now interact with the latest available wireless technology and can be used in a hospital environment. The two PDA devices used in this case study are the Palm Tungsten T3 and the Toshiba e800 Pocket PC.

Palm

The Tungsten T3 is selected for this study because of its flexible features. At the time of the study, the T3 was the most highly developed Palm device, with built-in Bluetooth wireless technology, 64 MB memory, and moderate screen resolution. To reduce screen scrolling, the T3 has a stretch screen display as well as portrait and landscape modes.

Toshiba e800 Pocket PC

The Toshiba e800 Pocket PC has both Bluetooth and Wi-Fi capabilities, with a memory capacity of 32MB. This has proven to be sufficient memory for the PDA-based medical programs. The e800 has a high-resolution VGA screen, 640 by 480; this improves the images on the Pocket PC and allows medical staff to read data with ease.

AVAILABLE WIRELESS TECHNOLOGIES

Wireless technology has been used for several years in the information technologies area; however, more recently there has been more fascinating development in wireless technology, allowing its useability in a wide range of areas. One such area is medicine: more and more new areas are being developed in which wireless technology is becoming a necessity (Bird, Zarum, & Renzi, 2001). Wireless technology in medicine allows mobility which provides greater advantage when it comes to patient care and patient diagnosis.

Bluetooth technology is an effective radio technology that allows for point-to-point or point-to-multipoint support which connects multiple Bluetooth devices. It has proven to be the most effective cable-free connection for wireless technology.

Some of the features of Bluetooth technology are: the transmission distance is real-world range between 10 and 42 meters, and transmission speed which provides throughput comparable to modems and DSL and security with Bluetooth integrates security mechanisms within several layers of its protocol. Connections are

Figure 2. The Toshiba e800 Pocket PC has the following features: Wi-Fi wireless technology—synchronisation, easy configurations for 802.11b wireless environment, 32 MB memory, portrait and landscape, VGA high-resolution screen/480×640 pixels (Toshiba, 2005)



generally configured to use authentication and 128-bit encryption. Applications can build their own security on top of Bluetooth connections to make communications even more secure; radio power output is as little as 1% of the power output of Wi-Fi (Palm One, 2004).

The 802.11x network is more commonly known as Wi-Fi, of which there are three general types: 802.11a, 802.11b, and 802.11g.

The most popular Wi-Fi is 802.11b, which has a throughput capacity of 11 Mbps and an approximate range of 50 meters. Due to bandwidth overheads and congestion, 802.11b will deliver data at about 7 Mbps, which is more than 10 times the performance of Bluetooth (Colin, 2002). The advantage of 802.11a and 802.11g is a five-fold increase in data rate over 802.11b (Colin, 2002).

These are only a few examples of the variety of wireless devices, some of which are being developed for specific use.

Medical Programs on PDA Devices

Apart from the Cerner Pocket Power Chart, other PDA medical software programs were installed on the PDA devices used in the study to assist in decision making and to increase efficiency. This allows medical staff members to access medical information at patient bedsides, reducing the physical load of books and decreasing search time. For example, on a ward round, a staff member may need to enquire about interaction between two drug types. There are two options in this situation: call the pharmacist, or look up the information at the nurse's station on the Internet or in a reference book. With the PDA at the patient bedside, the staff member now has access to the relevant information immediately. These medical programs used on the PDA were selected based on suggestions from medical staff members at the hospital.

iSilo is a program that converts file formats utilising compression for reducing document size. *iSilo* was used to reduce the size of the Australian Medical Handbook and the Westmead Medical Handbook so that they could be installed on the PDA device.

Medical Calculator is used in performing common calculations useful for clinical decisions or for analysing data in clinical decisions.

Medical Handbook is used by most junior medical staff as a resource to manage common medical and surgical problems.

MIMS (Monthly Index of Medical Specialities) provides medical staff and healthcare professionals with easy access to drug information required for making medical decisions.

Therapeutic Guidelines provides information to support decision making in the prescription of drugs.

HotSync allows programs to be uploaded to PDA devices from desktop computers.

Cerner Pocket Power Chart is a program that provides a mobile solution for hospital results reporting. It enables doctors and other healthcare providers to view patient results such as pathology and radiology reports.

The Cerner Pocket Power Chart has a username and password login as part of its program functions. Each doctor is allocated a username and password that are required for login to view patient lists. All hospital databases are secured behind firewalls, adding to the security of the device. The Cerner client software has built-in security—it is designed to erase all information on a PDA device if that information is not reviewed within 72 hours. This ensures that if a Palm device is lost or stolen, all Cerner Pocket Power Chart patient reports will be erased after 72 hours. This feature protects all patient information and upholds patient confidentiality.

Evaluation Research Method

To evaluate the useability and functionality of the PDA devices in a clinical environment, a survey was conducted. PDA devices were uploaded with medical support programs currently in use by some medical staff, such as MIMS, Therapeutic Guidelines, iSilo, MedCalc, the Westmead and Australian RMOs Handbooks, along with the current available version of Pocket Power Chart Palm OS client from Cerner Corporation, the vendor of the enterprise's results reporting system. The Citrix Client is used to view Cerner patient information and data on the Pocket PC handheld devices.

Doctors from various backgrounds were given PDA devices, complete with necessary programs, a user manual, and surveys to complete. They were allowed to work their way through the different programs in an unstructured way. After a week of using the PDA devices, observations were carried out to determine how often the programs and PDA devices were being used to assess patients and assist with decision making.

Survey

The survey was divided into three sections: the first section contained statements which were answered with agree, strongly agree, disagree, and strongly disagree. The second section consisted of yes and no answers, and the last section was a short, open answer section that allowed the medical staff to express their opinions about the PDA trial. The survey was completed after the trial period.

Statistical Analysis

A statistical analysis was carried out to calculate time utilisation in the clinical trial. Each medical staff member was timed before the

clinical trial and during the clinical trial. They were timed on their hospital duties on different hospital shifts, such as ward rounds, overnight shift, patient discharge, and new patient evaluation. This was carried out to determine how long it takes a medical staff member to carry out his or her duties with or without a PDA device. Times were noted and then put into a graph showing the range of time it took to complete, and the average time it will take to complete different duties in different hospital shifts (as outlined in the results).

Method Outline

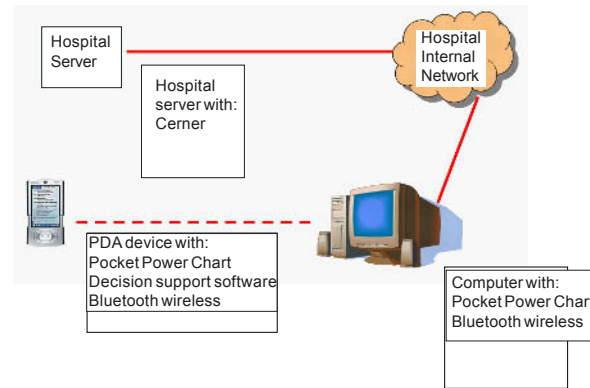
- A questionnaire was completed by the medical staff before the clinical trial. This determined general knowledge of PDA-type devices.
- Medical staff members were given a PDA device with a complete set of medical software and a user manual for a period of two to three weeks or more if required.
- After the trial period a survey was completed to determine how effective the medical staff found the PDA devices and the PDA-based medical programs.
- Observations were carried out to calculate the average time to access five patients with or without the PDA device.
- How frequently different medical programs were being used was also noted.
- The average time it took to prescribe medication with or without electronic medical databases or PDA-based programs was determined.

Synchronisation of Patient Information

HotSync

HotSync is a Palm-based computerised platform that allows data to be transferred between

Figure 3. Using HotSync and Bluetooth, wireless technology patient information on Cerner Power Pocket Chart is uploaded to the Palm Tungsten T3



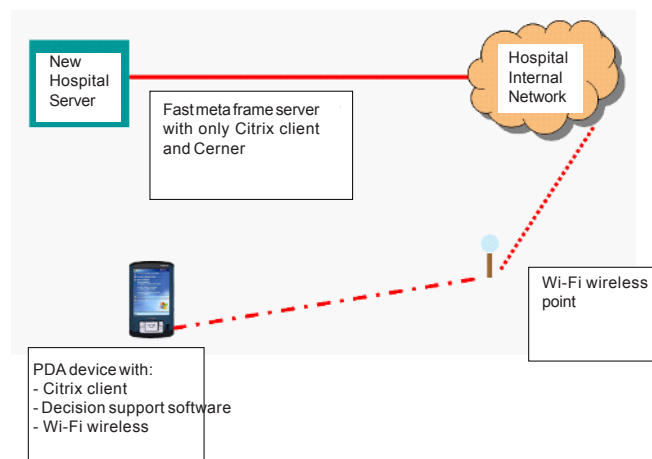
the Palm devices and the computer. This transfer of data is known as synchronisation and can be carried out through a cable connection or a wireless connection. Palm-based applications and programs are uploaded to the devices using HotSync synchronisation, as shown in Figure 3.

Citrix

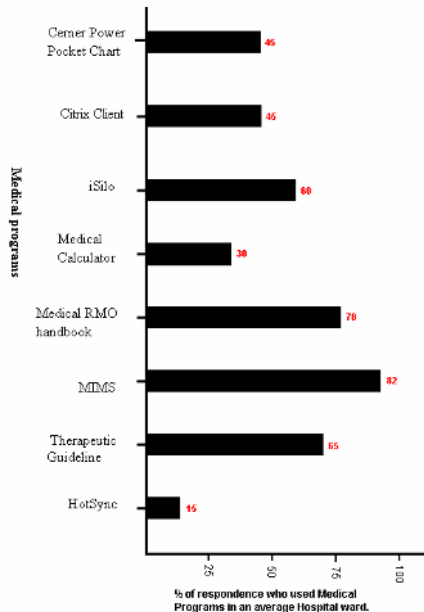
Citrix, a revolutionary software, delivers terminal services to a number of operating systems such as Windows, Linux, UNIX, and Macintosh.

In addition to these operating systems, Citrix also supports Palm and Pocket PC devices. A Citrix client allows a PDA to connect to a Citrix server, making it compatible to a standard desktop PC or laptop. As a result there are now practically no limitations to portability and compatibility of software restrictions on PDA devices. In addition, there is no longer a need to make a specialised application for both formats—Pocket PC and Palm. Rather, with the power of a terminal session through Citrix, the only requirements would be for the application

Figure 4. Through Citrix terminal sessioning and Wi-Fi wireless technology patient information on the Cerner program is captured on the Toshiba e800 Pocket PC



Graph 1. The percentage of respondents who used each medical Palm-based program on an average hospital ward (not to scale)



to run on a desktop such as a Windows operating system.

However, Citrix is not only a medium for reporting applications in terminal sessions, but also a client/server-based application that provides built-in layered security protocols allowing security and confidentiality. A Citrix client can be configured with a wide range of security provisions, including standard authentication using a username and password to extremely high levels of security using EAP (Extensible Authentication Protocol) (Citrix, 2004). Security controls such as certificates, smartcards, or even fingerprint and retinal scans can be used for identification and authentication. A Citrix client can also connect to a Citrix server via the Internet using the World Wide Web, allowing it to be published externally to the Internet, thus allowing ease of use and accessibility worldwide (a handy feature when you

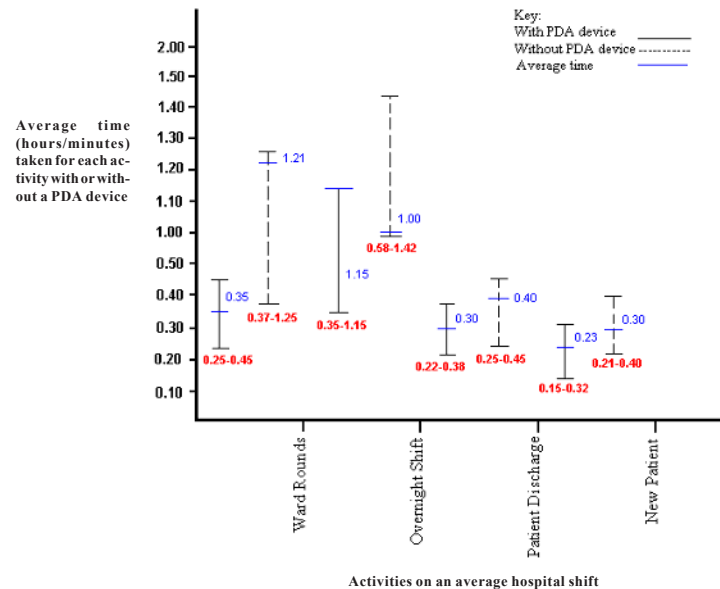
are out of the office and require a document stored on a network drive) (Citrix, 2004).

The Use of PDA-Based Medical Programs

Graph 1 outlines the percentage of PDA programs utilised during various hospital ward shifts. Cerner Pocket Power Chart (CPPC) was used by 45% of the doctors; the lower percent was due to a few problems with the CPPC program which includes the time (up to 24 hours) taken to upload and access new patient information and results on a PDA device. The usage may increase if the CPPC program is more efficient in uploading patient information and results. The Citrix client and CPPC had identical usage rates, as CPPC uses Citrix to access patient information. Furthermore, iSilo was used by 60% of the doctors. This was due to medical staff uploading other medical reference programs and using iSilo to convert them into PDA-based programs. The Medical Calculator was mainly used by specialist medical staff to convert and calculate test results when they were required, with approximately 30% reporting use of the application. The Medical RMO handbook was the second most popular program used by 78% of the doctors in a hospital shift. MIMS was the most used application, with 82% of the doctors each shift reporting that they used it, mainly for drug information and drug interactions. Therapeutic Guidelines was found to be the third most popular medical program used by 65% of the trial participants. HotSync had the lowest value of 15%, as its only role is to assist in uploading CPPC from the hospital Cerner desktop program. During an average shift, patient information is only refreshed if a new patient is admitted or new test results are required, meaning that HotSync was not required as often as the other programs.

The Evaluation of Wireless Devices Used by Staff at Westmead Hospital, Sydney

Graph 2. The range and average time to carry out medical tasks on ward rounds, overnight shift, patient discharge, and new patient assessment (not to scale)



Time Analysis of PDA Devices While Performing Medical Tasks

Graph 2 illustrates the varying timeframes required for medical activities with or without a PDA device. The graph outlines the average time range taken to carry out a task during different shifts at the hospital. It is evident from the graph that regardless of the hospital shift or task, the PDA devices and PDA-based pro-

grams assisted in reducing time taken to evaluate patients. For example, managing common medical problems and searching for medication for patients requires medical staff members to go back to the nurse station from the patient bedside to look up the necessary information. This can take up to 30 minutes as they go through several book references. But with the PDA devices, the medical staff can stay at the patient's bedside and search for the necessary information in less than a minute (this time is in

Table 1. Outlines the user percentage on a survey statement

Survey Questions	Percentage Outcome (%)			
	Strongly Agree	Agree	Strongly Disagree	Disagree
Did the PDA device improve access to patient information and necessary medical information at the patient's bedside?	23	58	5	14
The PDA device with its medical programs assisted in decision making and diagnosis.	18	78	0	4
Security level in this trial and on PDA application is sufficient.	15	44	12	29
Mobile devices such as PDA devices and mobile tailored medical programs can be seen as a daily part of a hospital environment.	21	59	9	11

accordance to observations made during the trial period).

Survey Outcome and Options

The survey conducted outlined the options of the users as well as the productivity of the PDA devices and medical applications and programs. Mobile medical programs and applications were seen as a convenient and essential part of daily duties around the hospital (Tschopp, Lovus, & Geissbuhler, 2002). A larger majority of the users considered the programs a fast and easy tool while making discussions and diagnosis, and approved its use and availability when required at the patient's bedside. The main concern was security of patient information in an event of a theft or lost property; this was secured with the 72-hour limitation on all patient information (Cerner program function) and a complete shutdown of the PDA device if power is not maintained which will erase all stored information in the memory. The security is handled with a user login and server firewalls which maintain the overall security of the hospital wireless system. It was noted that the Palm T3 added functionality of an expandable screen, reduced screen scrolling, and increased visibility. Also the VGA screen on the Pocket PC Toshiba e800 with high resolution provided x-ray images for reference. These increased the potential and further the use of such devices, making them suitable for medical operations.

CONCLUSION AND FUTURE DIRECTIONS

From the above results and feedback from clinical staff at Westmead Hospital, it can be concluded that such trial can be seen as a success and a route to possible implementation

of wireless technology. Similar clinical studies have looked at the possibility of PDA devices being introduced into hospitals for clinical use. Previous clinical trials of handheld devices have demonstrated both limitations and benefits to these devices. The PDA clinical trial had a few limitations; these were found in PDA based medical programs. Cerner Power Pocket Chart was found to provide only a limited amount of patient information and patient results. It limited patient results' viewing time to 72 hours and was unable to show previous results for comparison, a major disadvantage in clinical management. This reduced the use of CPPC, as medical staff had to go back to traditional methods to view complete patient details and results. For the CPPC program to be a successful tool in patient evaluation, improvement is required in areas such as patient administration details. However, other PDA medical programs combined together proved to be a very useful collection of reference tools for the medical staff. This also encouraged users to take the initiative and upload other PDA medical programs. To gain a better understanding of the benefits of PDA device usage, it may be important to extend the trial period to allow intermixing periods with and without the staff member having access to a PDA device. This will allow evaluators to measure the effects of PDA devices on hospital staff when the device is used for a longer period.

The main aim of this trial was to evaluate time taken to carry out different medical duties in the hospital setting (Ruland & Cornelia, 2002). Previous trials have concentrated on the functionality of the PDA device or opinion of the medical staff (Carroll, Tarczy-Hornoch, O'Reilly, & Christakis, 2004). This trial considered all three factors: staff option through the use of different medical programs and software, use ability of the PDA device in a hospital environment, and also accessed time during

different medical duties. As noted through the results, it is evident that the trial has proven to be successful in determining how effective this method of patient assessment can be, with the correct software environments.

ACKNOWLEDGMENTS

The author gratefully acknowledges Chris Liddle of the University of Sydney, the Westmead Hospital Clinical Pharmacology Department, the Westmead Hospital Information Technology Department, the Westmead Hospital Medical Staff, Citrix, Cisco, Palm One, Toshiba, and Therapeutic Guidelines.

REFERENCES

- Bird, S., Zarum, R., & Renzi, F. (2001). Emergency medical resident patient care documentation using a handheld computer devices. *Academic Emergency Medicine*, 8(12), 1200-1203.
- Carroll, A., & Cristakis, D. (2004). Pediatricians' use of and attitudes about personal digital assistants. *Pediatrics*, 113(2), 238-242.
- Carroll, A., Tarczy-Hornoch, P., O'Reilly, E., & Christakis, D. (2004). The effects of point-of-care personal digital assistance use on resident documentation discrepancies. *Pediatrics*, 113(3), 450-454.
- Citrix. (2004). Retrieved June 24, 2004, from <http://www.citrix.com/lang/English/home.asp?source=google&keyword=citrix&sitrackingid=399741>
- Colin, D. (2002). *Going wireless*. Retrieved August 1, 2004, from <http://www.projectorcentral.com/wi-fi.htm>
- Palm One. (2004). *Bluetooth wireless technology*. Retrieved March 11, 2004, from <http://www.palmos.com/dev/tech/bluetooth/>
- Palm One. (2005). *Palm One Tungsten T3 image*. Retrieved February 10, 2005, from <http://www.palmone.com/us/support/tungstent3/>
- Ruland, C., & Cornelia. (2002). Evaluating a support system for preference based care planning at the bedside. *Journal of American Medical Informatics Association*, 9(2), 192-201.
- Toshiba. (2005). *Toshiba e800 Pocket PC image*. Retrieved February 10, 2005, from <http://www.isd.toshiba.com.au/cgi-bin/>
- Tschopp, M., Lovus, C., & Geissbuhler, A. (2002). Understanding usage patterns of handheld computers in clinical practice. *American Medical Informatics Association*, 1531(605), 806-809.

Chapter VIII

Mobile Comprehensive Emergency System

Samir El-Masri

University of Western Sydney, Australia

ABSTRACT

This chapter is an application of Mobile Web Services into the health sector and specifically to the emergency system where the communication between a number of parties is critical in terms of time, efficiency and errors. The proposed application is to implement a mobile system based on cellular phone network in ambulances. It will equip doctors with mobile devices that have the capability to be connected to the Internet network with a bandwidth that makes it feasible for the doctor and the ambulance to access the health record of a patient from the database and to communicate with hospitals with enough speed. This chapter will illustrate the new proposed Mobile Comprehensive Emergency System (MCES) application that is based on Web services provided by static and mobile servers. The implementation of this new system will enhance the current system communication and make it more reliable, consistent, quick and free of human errors.

INTRODUCTION

Most of the existing communication systems set to communicate between ambulances and hospitals rely on radio communications. Some new systems include a computer system as a tool to help in the management of communications like the Computer-Aided Dispatch (CAD), implemented in 1995 in Victoria, Australia. This system was enhanced with a Medical Priority

Consultant's Advanced Medical Priority Dispatch. The computer version of the new assistance was introduced in April 1998. The Victorian system is considered one of the best emergency systems in the world in providing clinical information about the patient to the hospital and recommending some care from the hospital to the ambulance. The system is also backed by an automatic vehicle location system to locate the ambulances (Service, 2002). Another ad-

vanced system is Hospital & Emergency Ambulance Link (HEAL), implemented in Singapore. The HEAL system is based on wireless data communication between ambulances and hospitals, and it assists hospitals and doctors at the emergency departments of the hospitals with information about patients provided by ambulances. HEAL also assists ambulances with medication recommendations from doctors. This system has been built on the public wireless network and is based on server-client architecture (Ministry of Health, National Computer Board, Singapore Civil Defence, & Singapore General Hospital, 1998).

The proposed new system is more comprehensive than any existing emergency system in terms of the number of parties involved, and it is more advanced in terms of technology proposed. The new system is also intelligent when it comes to finding the right ambulance, hospital, and doctor that are suitable for the conditions and location of the accident. The main advantage and strength of the new system comes from the Mobile Web Services technology that would be used in the system (Chatterjee & Webber, 2004; El-Masri, 2005). This technology can overcome any problems of interoperability between systems running different applications based on different programming languages on different platforms (El-Masri & Suleiman, 2005; El-Masri & Unhelkar, 2005). More details of the new system will be found in next sections of this chapter.

APPLICATION OF MOBILE WEB SERVICES TO A COMPREHENSIVE EMERGENCY SYSTEM

In this section, the new system business process will be explained in detail, followed by the

technical details about the new technology applied.

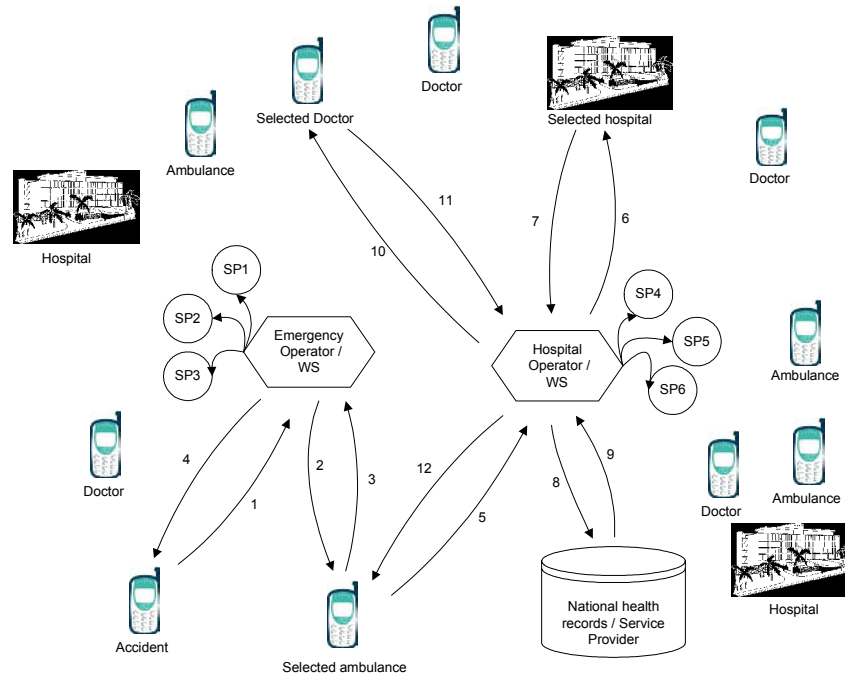
Case Study

In the case of a car accident, for instance, and when we dial 000 (in Australia) for emergency and ask for an ambulance, the current operator (human) on a phone-based system tries to find the nearest available ambulance to be sent to the accident location. After picking up the patient, the ambulance officer heads for the nearest and available, suitable hospital.

In this chapter, the proposed system is a comprehensive emergency system based on Mobile Web Services. As shown in Figure 1, when there is a car accident, any nearby mobile holder can enter information about the accident such as how many cars are involved, how many people are injured, and how far the caller is from the accident. The mobile phone (First Informer) will send this information to an operator, which is a Web services centre. The operator will access the situation based on the incoming information from the First Informer Mobile Phone (FIMP).

Let us imagine the following scenario in which police, the fire brigade, and an ambulance are needed. The Web services-based operator (emergency operator/WS) will look through its service directory, which is timely, updated based on the mobile system locator, to find the nearest available police car, the nearest available fire brigade, and the nearest available ambulance. The operator will send electronic emergency requests to the selected police car, fire brigade, and ambulance, and request confirmation from their systems. The systems of the police car, fire brigade, and ambulance can be based on a mobile or wireless system. To simplify the case study, only the communication with the ambulance will be explained. This communication scenario is applicable to the police and fire vehicles as well.

Figure 1. Comprehensive emergency system



After personal intervention from the ambulance officer, the system can send confirmation that the ambulance is heading for the accident location. The operator (a proxy server that processes services from providers and sends information back to the client application) will calculate the distance and the time needed for the ambulance to reach the accident place; this information can be transferred to the FIMP through the operator. In case of unavailability for unknown reasons, the operator can look down the list for the second available ambulance. The same scenario is applicable to the police car and the fire brigade. Once the ambulance has reached the accident and its officers assisted the patient or the injured, they will input into their mobile system the patient's conditions. This information will be electronically sent to a hospital operator (Web services

server), which instructs the ambulance to go to the suitable hospital (nearest hospital, specialised in the patient injuries with available beds and doctors) and passes on the patient's conditions to that hospital.

At the same time, doctors and nurses at the emergency department of the selected hospital would prepare for the coming patient and they would know from the system the expected time of arrival. The hospital system and the ambulance can also access the patient health record to be passed on to the doctors. If there are no doctors available within the hospital to treat the coming patient, the hospital will search the operator for the right doctors outside the hospital through their mobile systems, and the doctor can receive information about the patient from the ambulance. After accessing the patient health record on his/her PDA, the doctor can

recommend to the ambulance officers some medications to be given to the patient before reaching the hospital. The doctor can also discuss with other doctors the case of the patient. The emergency department will be informed about the time the doctor can reach the hospital. The ambulance system and the patient health record also will automatically assist the officers in giving the right medications and avoid some allergy problems or history diseases.

SERVICE CONSUMERS AND PROVIDERS

In Figure 1, both operators play the role of Web services registerer, where the service provider and the service consumer are registered and authorised to use the system. The accident mobile or the FIMP is a service consumer that uses the unique available service to report the accident; it can also request a response when FIMP is part of the accident. FIMP is a service provider also, as it provides the emergency operator with its location automatically. The selected hospital will be a service provider when it provides its location to the emergency operator. The selected ambulance will also use a client application to consume services provided to the hospital operator by hospitals to determine the right hospital. The ambulance mobile system provides services to the hospital and doctor when it transfers information about the patient's conditions. The selected hospital system provides services about its availability to the ambulance. The doctor system has a mobile client application to consume services provided by the national health records system and hospitals.

BENEFITS OF THE PROPOSED SYSTEM

The following are the advantages the new comprehensive emergency system can have over the current existing emergency systems. The new system:

- represents an application of a new technology (Mobile Web Services);
- presents a complete computerised system based on MWS;
- is completely based on the new wireless cellular network (it should use at least 3G or 4G technology; see the chapter on understanding Mobile Web Services and their role in integrating mobile devices); and
- is based on an intelligent system in terms of deciding on selecting the ambulances, hospitals, and doctors.

AVAILABLE TECHNIQUES

Currently, a few tools are available to build mobile applications on mobile devices to consume Mobile Web Services. The most popular available tool is the Microsoft Mobile Internet Toolkit, which is an additional part of Microsoft

Figure 2. The new mobile Web application project from Visual Studio .NET

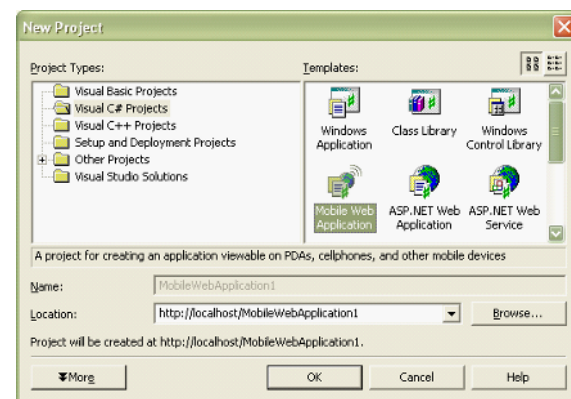
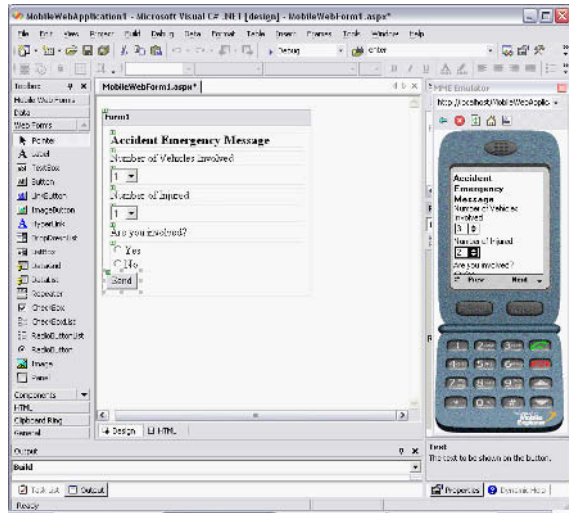


Figure 3. Building a mobile Web application to report an accident from your mobile device



Visual Studio .NET. Other tools like Java or IBM are also available.

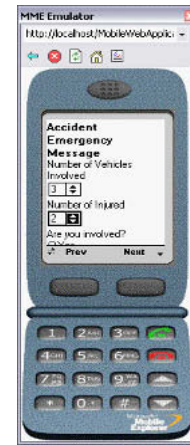
THE MICROSOFT MOBILE INTERNET TOOLKIT

Microsoft, using .NET framework and Visual Studio .NET, makes it simple to build Web services and develop mobile client applications to consume them. In order to be able to develop a mobile application that can consume WS, a free package called “Microsoft Mobile Internet Toolkit” needs to be downloaded and installed from Microsoft.com. Figure 2 shows the start page of the Visual Studio .NET where you can create a new project, select the language you want to use, and the appropriate template.

Figure 2 shows that C# has been selected as a programming language and “Mobile Web Application” has been selected as a template to build a mobile Web application.

Figure 3 shows a simple mobile application built to report an accident. It includes two

Figure 4. Mobile device emulator of FIMP

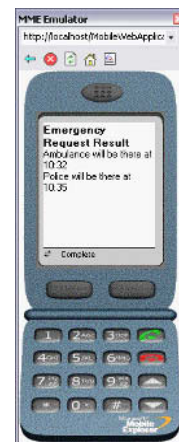


selection list controls to select the number of vehicles involved and the number of injuries, and whether or not you are involved in the accident.

Figure 3 shows a mobile phone emulator on which the application is displayed as it would be implemented into real mobile devices.

A proposed FIMP mobile application appears in an emulator, as shown in Figure 4. The application was built to suit a mobile phone,

Figure 5. The received results from the emergency operator at FIMP



PDA, or any other mobile device, as it is simple, small, and easy to access, and there is no need to enter any data; the user needs only to select the available data.

Figure 4 demonstrates that FIMP indicates that there are three vehicles involved and there are two injuries as a result of the accident. FIMP has also selected the option as it is involved in the accident; this way FIMP will receive a report back from the operator about the estimated time the ambulance needs to reach the accident location.

Once the operator has received the message about the accident, and found the right ambulance and police car, it will report back to FIMP if it is involved in the accident. Figure 5 shows the report and the estimated time needed for the ambulance and the police to be on the scene of the accident. It shows the ambulance will reach the accident location at 10:32 and the police at 10:35. It should be noted that there is no need for the fire brigade to be dispatched; this is estimated automatically by the system.

It should be noted that .NET framework does not provide tools to build services to be provided on mobile devices. This issue is outside the scope of this chapter; a new proposed system for that purpose is available in El-Masri and Suleiman (2005).

CONCLUSION

This chapter proposes a complete and comprehensive emergency and medical assistance system based on Web services and Mobile Web Services. In this system, mobile devices have been used to support mobile services as service providers and mobile client applications capable of consuming Web services and mobile Web services. The implementation of this system will dramatically reduce the price and the time of current communication systems be-

tween all parties mentioned above, and the most important benefit is the automatic search carried out for the right and available service at a reduced time. The accuracy is also one of the great benefits the system can offer. What we have not looked at closely is the security of the system and particularly the authorization to access the public health records. The basic security to be adopted by the system is the implemented authentication in the mobile device by a subscriber identity module (SIM) card used in a GSM system. More security measures for the whole system will be further investigated.

REFERENCES

- Chatterjee, S., & Webber, J. (2004). *Developing enterprise Web services*. Englewood Cliffs, NJ: Prentice-Hall.
- El-Masri, S. (2005, September 26-28). *Mobile comprehensive emergency system. Proceedings of the 2nd International Conference on Innovations in Information Technology*, Dubai, UAE (pp. 125-131).
- El-Masri, S., & Suleiman, B. (2005, September 26-28). Providing Web services on mobile devices. *Proceedings of the 2nd International Conference on Innovations in Information Technology*, Dubai, UAE (pp. 132-141).
- El-Masri, S., & Unhelkar, B. (2005, May 23-26). Modelling XML and Web services messages with UML. *Proceedings of the Information Resources Management Association 16th International Conference*, San Diego, CA (pp. 250-258).
- Ministry of Health, National Computer Board, Singapore Civil Defence, & Singapore General Hospital. (2005). *Launch of Hospital & Emergency Ambulance Link (HEAL) pilot project*.

Retrieved March 13, 2005, from <http://www.ida.gov.sg/idaweb/media/infopage.jsp?infoagecategory=ncbarchivemediareleases.mr:media&versionid=4&infopageid=I1013>

Service, M. A. (2005). *The Metropolitan ambulance service*. Retrieved March 13, 2005, from <http://www.ambulance-vic.com.au/opservices/communications.html>

Chapter IX

Application of Mobile Technologies in Healthcare Diagnostics and Administration

Ketan Vanjara
Microsoft Corporation, India

ABSTRACT

This chapter explores various advancements in mobile devices and corresponding software applications that enhance diagnostics and administration in the healthcare domain. This chapter further proposes networking and integration of these devices with the existing networks and devices as further development in healthcare.

A ROAD ACCIDENT SCENARIO

In 1992, a colleague of mine was traveling from our offices in western India to our manufacturing facilities about 50 kilometers away. He had worked late the previous night and again come to the office early in the morning. We had our

lunch together and he left for the manufacturing facilities (factory) around noon, driving his own car. Even after two hours, he had not reached our factory and, increasingly, the office and the factory administrators were panicking, trying to locate this employee of ours. Around 6:30 p.m. we received a call from the general hospital of

the city that a semi-conscious accident patient had been admitted to the hospital. In short, my colleague had met with an accident around 1 p.m., and he was admitted to the hospital only around 6 p.m. The treatment could not begin before 8 p.m. The end result was that he became paraplegic. Based on the information provided by the doctor on duty, had he been treated within three hours of the accident, paraplegia could have been avoided. Did mobile devices—or their lack of—have a role to play in this incident? The answer has to be a ‘yes’.

Though just a little over a decade ago, this era was devoid of any mobile devices in India. My colleague did not have a cell phone, nor did his family (it took a while to contact them). His car did not have a Global Positioning System (GPS). The hospital and doctors too did not have any mobile devices. Finally, even after arriving at the hospital, the doctors had to wait for the family before starting treatment, as they needed medical information of the patient such as existing diseases, allergies, sensitivities, and so on. Overall, the above scenario shows the lack of our ability to manage emergency medical situations without immediate communication. Today, though, with the advancement of mobile technologies, such events are better managed. However, there is still a need to provide integration of various mobile tools with the existing devices, technologies, and networks in order to reap full benefits of mobility in the healthcare sector in general and emergency medical situations in particular.

This chapter explains how mobile technologies are shaping the face of healthcare today. This chapter also discusses the potential for improved healthcare through mobility in the future. The primary focus of this chapter is on the integration of mobile devices/gadgets in diagnostics, healthcare administration, and healthcare information systems. Finally, this

chapter also proceeds to identify issues and challenges involved in such integrations and how to handle them.

THE HEALTHCARE LANDSCAPE

Healthcare is one of the major domains in life that is full of amazing advancements. However, it also comes with its ever evolving challenges. For example, of late this domain has been mired in controversies on ethical issues like cloning and stem cell research. While medical science still continues to struggle with challenges like eradication of malaria and treatment of cancer and AIDS, its achievements have not been mean by any standards.

The progress made in the last 20 years (is) amazing if one looks at it now. In short, the advances made in the last 10-15 years are equal or even surpass the advances made in the last 100 years. (Majeed, 2005)

Medical science has discovered new diseases and invented new treatments and therapies. In fact, a number of medical scientific disciplines like Pharmacogenomics that allows creation of personalized medicine, toxicokinetics, proteomics, therapeutic vaccines, stem cell research, bioinformatics, and telemedicine, to name a few, evolved in the last 20 years.

Complex diagnostics, short-duration intensive care, acute medicines, and micro-surgeries are all becoming common in today's healthcare.

Another major contributor to advancements in medical sciences has been the developments in digital technologies. Be it bioinformatics or telemedicine, the need for modern healthcare to offer the best treatment for complex diseases in the fastest possible manner with resource constraints makes it highly dependent

on digital technologies. Data-intensive areas like bioinformatics also need to depend heavily on high-end computing devices.

Thus, today's healthcare is characterized by the following:

- **Multiplicity and Complexity of Diseases and Treatments:** The breadth and depth of knowledge on various medical disciplines, diseases, their symptoms, causes, and various alternatives of treatment have been growing exponentially in the last couple of decades. This has led to increased complexity in the fields of diagnostics and therapies. So much has been the growth in knowledge and complexity that it is almost impossible for a medical practitioner to function effectively without supporting devices like computers and PDAs.
An indicator of this fact is the vast amount of literature available on all the medical disciplines—a simple search on Google (2005) returns thousands of medical journals, 1,440 of which are free. Hundreds of viruses and their mutations have been identified for something as simple as the common cold (Health-Cares, 2005).
- **Quick Response:** Given that the number of accidents is increasing every day, the demand for quick response from medical services is growing too. Continuous efforts are on to reduce the interval between the time a person needs medical attention to the time when treatment is provided to her. This has led to provisioning of preliminary treatment as soon as the patient is contacted—at the accident site, at the patient's home, or even in the ambulance taking the patient to a clinic or hospital. Mobile clinics and surgery ambulances are quite common nowadays to meet this need of quick response.
- **Quick Cure:** Quick response alone is not enough. A typical patient today expects to get cured much faster so that he can get back to work and his normal life. This need drives the entire industry, especially pharmaceutical companies, to constantly research and produce better drugs and come up with drug delivery systems that offer faster cure (Howard, 2004).
- **Mobility and Remote Assistance:** Mankind is more mobile today than it has ever been before. This increase in mobility has given rise to the need for remote medical assistance. Despite their mobility, a patient and a doctor need to stay in touch with each other to get/provide remote assistance. Such remote assistance can take various forms, ranging from simple things like monitoring a patient's progress and response to treatment, to more complex and technology-intensive issues like telemedicine or tele-surgery.
- **Cost-Price Paradox:** While it is quite obvious that the emerging nature of medical services as well as demands on them result in increased costs, there is constant pressure to drive the price of healthcare southwards. Besides providing better value for money to patients, reduced price also makes healthcare affordable to many more people. Thus, it is imperative upon various players in the healthcare sector to provide superior and continuously improving products and services at continuously reducing costs. This requires a lot of automation and innovation.
- **Resource Constraint:** Increase in multiplicity and complexity of medical disciplines, diseases, and therapies requires specialized skills. Thus, we have specialists in every branch and sub-branch of medicine. By their very nature, specialist skills are scarce. However, the demand

on such specialist skills keeps on rising. This creates a resource constraint and a need to be able to utilize the same resources effectively across different locations or geographies.

- **Data Driven:** Given the plethora of variables that need to be considered and evaluated to provide effective treatment, the medical profession has become highly data driven. Besides supporting research, such data actually influences, if not drives, treatment. Such dependence on volumes of data necessitates use of digital devices.
- **Return of Alternative Therapies:** These include medical disciplines like the age old ayurveda, naturopathy, homeopathy, and so on. While their acceptance is growing among many people all over the globe, technology has not made as many inroads into these branches of medicine as others. This presents a huge opportunity. We shall discuss one such potential in this chapter.

THE MOBILE TECHNOLOGY LANDSCAPE

After surveying the healthcare landscape, let us take a look at the mobile technology landscape as well. The handheld devices of today, like cell phones and PDAs, are far more powerful than the large supercomputers of yesteryears that occupied large space. In addition to this shrinking of sizes, there are quite a few other significant developments taking place in this arena. Some of the recent developments that help us understand the mobile technology landscape are as follows:

- **Nanotechnology—Miniaturization Shift:** Handheld devices and various other gadgets related to mobile technologies are getting smaller and smaller. The scale is

moving from micro to nano. Nanotechnology, considered to be one of the foundations of future technology revolutions, enables many unthinkable things like storing several gigabytes of data on a millipede nanodrive of the size of a postage stamp (Global Change, 2005). We already get 60 GB of storage in an iPod (<http://www.apple.com/ipod/>). PDA PC watches and pocket-sized holographic projectors are on their way to the market. One of the constraints of adopting mobile technologies has been smaller output interfaces in the form of display screens. Hologram projectors resolve this problem effectively (<http://www.lightblueoptics.com/index.html>; Howard, 2004).

- **Software Power:** While hardware devices are offering more (functionality) in less (smaller devices), software is not to be left behind. There are thousands of applications created for these small devices that have larger functionality and very small footprint. Be it Win CE operating system, various MS Office applications, or entertainment software like iTunes—all of them provide extensive functionality while occupying much less space on the devices (Microsoft, 2005).
- **Convergence:** The world talked a lot about convergence of information and communication technologies in late nineties. However, such convergence is an ongoing process that is still continuing. What we are seeing today is convergence of various advanced technologies like computing, mobile telephony (cell phones), and connectivity (Internet) on a single handheld device like the HP iPAQ 6315 (2005) Pocket PC—Phone Edition. Pretty soon, positioning (GPS) too will be integrated. Handheld GPS receivers are already available in the market. Microsoft's product *Streets & Trips 2005* with

handheld GPS locator combines Streets & Trips software with a GPS plug-n-play receiver hardware component, to create a planning and driving tool that turns a laptop into a real-time GPS tracking device (Microsoft Streets & Trips 2005 GPS, 2005). Very soon we may see all these features available on nano devices, resulting in the handheld devices having more processing power, superior features, larger storage, longer battery life, and becoming more universally accessible than the tablet PC of today. Such convergence of technologies will empower and enable users to work far more efficiently and effectively than ever, and from wherever.

- **Natural Interfaces.** Another major area of advancement in technologies is natural interfaces. These include voice recognition, handwriting recognition, and visual recognition. One of the greatest hindrances to adoption of mobile devices has been the input interfaces in the form of the keyboard or mouse. The smaller keypads of handheld devices make it difficult to input large content. Voice and handwriting recognition address this issue. Visual recognition addresses issues related to security. Very soon we will also see interactive surfaces like floors, tables, walls, and so on (<http://www.naturalinteraction.org>, <http://www.hcirn.com/>, <http://nooface.net>).
- **Function-Specific Handheld or Nano Devices.** While information and communication technology industries are creating newer, smaller, and smarter versions of technology-driven handheld devices like cell phones, pocket PCs, and PDAs, industries like healthcare are creating and continuously improving functionality-driven handheld devices to perform specific functions related to healthcare. These include diagnostic devices like Smart Pill

from Smart Pill Diagnostics and Amplichip from Roche (Medgadget, 2005a).

- **Mobility and Wireless:** With mobility comes the need of wireless. One cannot be truly mobile with wired devices. In addition to wireless devices, we are also seeing the creation of connectivity infrastructure in the form of cellular and Wi-Fi networks.

A cellular radio network is a radio network made up of a number of radio cells (or just cells) each served by a fixed transmitter, normally known as a base station. These cells are used to cover different areas in order to provide radio coverage over a wider area than the area of one cell. (Cellular Network, 2005)

Wi-Fi networks (also known as wireless LAN) use radio technologies called IEEE 802.11b or 802.11a to provide a secure, fast, and reliable wireless connection. Wi-Fi functions through a transmitting antenna which is usually linked to a DSL or high-speed land-based Internet connection and uses radio waves to beam signals. Another antenna, which is in the laptop or PC, catches the signal. The signal, usually, has a range of about 300 feet for most home connections. The farther the user is from the signal, the slower the connection speed. (Shaheen, 2005)

Our cell phones work on a cellular network, while mobile computers use various other wireless technologies to network, like infrared, Wi-Fi, radio-frequency, Bluetooth, and so on. Recently, cell phones that can work on both cellular and Wi-Fi networks, and also seamlessly switch between the two, have also arrived (Computerworld, 2004).

- **Cost Paradox:** The cost paradox in the technology landscape is quite opposite of

that in the healthcare space. In technology areas, the normal trend is to get more for less. From memory to storage to functionality—hardware or software—we either get the same product cheaper or a better product at the same price with passing time. This puts technology in an enviable position where it can be and it is complementary to all other industries or needs.

- **Business Intelligence:** Due to increasing specialization and complexity in various domains today, decision making in today's world is a function of a multitude of variables. This has necessitated creation of huge databases. These databases need to be analyzed almost real time to support decision making. Data mining and business intelligence come in handy here.

APPLYING MOBILE TECHNOLOGIES IN HEALTHCARE

Having surveyed the healthcare and mobile technologies landscape, we now move on to explore the possibilities of convergence of both. Though the possibilities are endless, we will keep our exploration limited to:

- **Diagnostics:** With the exception of certain alternate medical sciences, treatment of therapy nowadays is seldom provided without diagnosing the symptoms through a plethora of tests in the areas of pathology and radiology, as well as physical examination. The field of diagnostics, therefore, offers many challenges—in other words opportunities—for mobile technologies.
- **Healthcare Administration:** Healthcare administration includes the complete cycle

of healthcare provisioning from the moment a patient seeks the service until the patient is cured or no more needs the service. This includes activities like consulting, symptom recording, diagnostics, analysis and interpretation of test reports, prescription of treatment, therapy or surgery, drug delivery, post-operative support, and so on. Due to various characteristics of the healthcare sector mentioned earlier, mobile devices can make a lot of contribution in this area as well.

- **Healthcare Information Management Systems:** This includes creation, maintenance, and global provisioning of all healthcare-related information like medical records, test reports, receipts, payments, service levels, and so on. Not all information is critical such that it needs to be provided anytime, anywhere. However, information like medical records, test reports, service levels, and so forth do need to be universally available.

MOBILE TECHNOLOGIES IN DIAGNOSTICS

Among all the applications of mobile technologies in the healthcare space, diagnostics is the most challenging as it involves integration of multiple technologies. A patient has to go through a series of diagnostic tests to enable a doctor to provide appropriate treatment. There are two major bottlenecks around this: (1) in urban areas, where the diagnostic labs are normally located, the patient population is very mobile and also adversely affected by traffic conditions of modern urban areas; and (2) in rural areas, where the patients are not that mobile and rarely have a diagnostic lab. Mobile technologies can resolve both these problems.

In addition, mobile technologies can also assist in cases of emergencies, like accidents or elderly patients that cannot travel much.

Conventional diagnostics can be broadly classified into four areas: pathology, radiology, scans (MRI, CT, etc.), and pulse (ECG, physical pulse examination, etc.).

The good news is that today we have miniature devices that are useful in all types of diagnostics. In the beginning, personal wellness monitoring meant the bathroom scale and the mercury thermometer. Then they went electronic, and home blood-pressure devices became common. Today, it would not be shocking to see a home cardiac defibrillator sitting in the corner (HealthDay, 2005). However, only some of the diagnostics devices have become popular so far. Others are at various stages of development, clinical trial, or market deployment.

Let us briefly look at some of the handheld or miniature devices for diagnostics that are already being used or about to be deployed in the healthcare market. A brief description is given on the functionality of each device (<http://www.medgadget.com/archives/diagnostics/index.html>).

- **Blood Glucose Monitor Systems/Diabetes Systems:** These devices help in monitoring/determining the glucose levels in blood. They have become quite sophisticated over a period of time, and the latest models have many features like no strip handling, 250-test memory, and so on.
- **Wrist BP Monitors:** They look very similar to electronic wrist watches and are used in determining the systolic and diastolic blood pressure of a patient. The advanced versions can store up to 90 readings and are very lightweight, operating on AA or AAA batteries.
- **Oximeter:** These devices determine the degree to which hemoglobin on the red blood cell is loaded with oxygen (“oxygen saturation”). Sophisticated models have additional features like measurement of pulse rate and SpO₂, verification of endotracheal tube placement, and changes in carbon dioxide level with the help of a CO₂ sensor.
- **Automatic External Defibrillators/Pacemakers/CHF Monitor/Fluid Status Monitoring:** These are electrical devices used to counteract fibrillation of the heart muscle and restore normal heart-beat by applying a brief electric shock. In addition, they have built-in monitors that track the problem to prevent crisis. The fluid status monitoring device sits under the patient’s left collarbone and sends out electrical impulses to monitor liquid levels; whenever the level is near critical, the device alerts the patient to get help before the problem turns critical. Doctors also can check the monitors by phone, using computerized hookups in their patients’ homes.
- **iPod:** The iPod is not just for music any more. Radiologists in the United States, Europe, and Australia are now using iPod devices to store medical images. “This is what we call using off the shelf, consumer market technology,” says Osman Ratib, MD, PhD, professor and vice-chairman of radiologic services at UCLA. “Technology coming from the consumer market is changing the way we do things in the radiology department” (Ratib, 2004). A software (OsiriX) has also been developed to automatically recognize and search for medical images on the iPod. When it detects the images, they automatically appear on the list of image data avail-

able—similar to the way music files are accessible by the iTunes music application.

- **Portable Patient Monitors:** These devices monitor and provide digital display of mean and diastolic blood pressures, heart rate and respiratory rate, temperature (a, b, or differential A-B) based on two temperature inputs, blood pressure, and ECG trace. Several alarms can also be set in these monitors based on over/under pulse rate, pressure, or temperature (<http://www.pemed.com/pmonitor/pmon.htm>).
- **GI Monitoring System:** A patient swallows this slightly larger-than-multi-vitamin-sized capsule containing two subminiature radio transmitters. The capsule's transmissions are received by a small, patient-worn mobile receiver/controller. Site-specific data is captured in real time as the capsule passes through the GI tract. It also measures pH, temperature, and pressure. The capsule is not absorbed, nor does it interact with the GI tract in any way other than its propulsion via peristalsis, eventually exiting through the colon. The future-generation systems are expected to enable real-time, site-specific tracking of the capsule's position within the GI tract, transmit pictures, and also facilitate drug delivery.
- **Handheld MRI:** The primary reason for the large size of conventional MRI scanners is the use of large magnets that have to be chilled by complicated cooling systems. Igor Savukov and Michael Romalis of Princeton University have shown that a device called an atomic magnetometer can detect magnetic signals from water without giant magnets or complicated cooling systems. This has given rise to a distinct possibility of handheld MRI scanners in the future that would image tissues inside the body as easily as a digital camera takes a photo.
- **Wireless/Handheld X-Ray Sensor:** A company has developed a wireless direct digital radiography system for dental x-ray examinations that communicates images it captures with a dentist's computer. There are quite a few models of handheld or portable x-ray instruments available now. They provide greater flexibility, are optimal for remote uses, capture high-quality images, transfer them digitally to computers, and reduce radiation exposure.
- **Oral Swab Kit:** Various organizations are prototyping oral swab kits or similar devices for saliva-based diagnostics that can detect exposure to a variety of substances, from narcotics to anthrax to common bacteria and viruses including those of HIV. These will be quantum enhancements to rapid saliva tests that are already in use. These devices would increase ease of detection and accelerate response time whether they are used in the middle of a public health incident or in a busy doctor's office. The devices are expected to communicate captured data in digital form as
- **Wireless Miniature Devices for CHF Patients:** These are technologies integrated into minute implants, requiring no antenna, wires, or connecting leads, that allow a tiny device implanted deep inside the body to communicate wirelessly with other implanted devices and external systems. Such devices include a device that offers on-demand, non-invasive means to monitor intra-aneurysm pressures following endo-vascular graft procedures and a device for measurement of pulmonary artery pressure, which is the most important hemodynamic indicator in heart failure. A one-time, minimally invasive catheter-based procedure allows unlimited,

non-invasive home-based monitoring of a CHF patient's hemodynamic status.

- **Other Handheld Wireless Diagnostic Devices:** These include holographic sensors, PathoTester (a wireless handheld system that can detect growth of microbial pathogens using nanocartridge-based technology that contains microbe-specific sensor holograms), non-invasive glucometer (that measures blood sugar in the body based on the amount of heat released from the tip of the finger—metabolic heat conformation), Glucose Monitoring Watch (which continuously displays blood glucose through use of photo acoustic waves originating in the blood vessels), and Wireless Objective Hearing Assessment Systems.
- **Pulse Meter for Ayurveda:** Ayurveda, the science of life and longevity, is an ancient Indian system of holistic medicine, which has re-emerged as an important form of holistic therapy throughout the world. Its essence is the conjunction of the body and the mind, and their coordination that achieves optimal health and happiness. This science is heavily dependent on pulse examination for diagnostics *Nadi Pariksha*. This involves determining *Prakriti* and *Vikriti* in pulse, understanding the seven levels of pulse, and integrating qualities of the pulse with signs and symptoms in clinical assessment. Future enhancements of pulse meters are expected to fulfill the needs of Ayurveda as well.

There are tele-healthcare gadgets galore, and the list can go on and on. Take for instance an all-in-one device that helps patients do the following: measure blood pressure, blood sugar, blood oxygen (SpO₂), measure temperature and weight, record peak flow, record stethoscope sounds, and take ECGs.

As of now many of these devices are not network ready for communication with other wireless devices like cell phones or PDAs, nor are all of them following any common communication standard. However, very soon they will be getting linked wirelessly to home computers and cell phones, as a health revolution is creeping quietly into our lives.

Where it is going to take us is anyone's guess, just as whether new age health monitoring will make us live longer. But the IBM Corp., for example, has developed a little electronic pillbox that may come in handy as the Baby Boomers edge into the years of forgetfulness. The box sends a signal to a mobile phone every time a pill is removed. If patients forget a pill, or take too many pills from the box, they get a friendly reminder phone call.

For patients who need frequent monitoring of vital signs, an IBM wristband device measures blood pressure and heart rate at the push of a button, and these are transmitted automatically to medical personnel.

Sun Microsystems and MedicTouch have introduced the Pulse Meter, which they call the 'ideal health monitoring solution for sport enthusiasts, the elderly, rehabilitation outpatients and healthcare providers; providing health-monitoring anytime, anywhere.' A Pulse Meter user connects the sensor to a hand, starts the program on a mobile phone and the pulse is displayed within seconds on the phone screen, archived, customized for user, and transmitted. (HealthDay, 2005)

Medical monitoring and treatment are going wireless and Internet based, both at home and at the hospital. HealthDay (2005) reports about tele-healthcare gadgets designed for initial, basic in-home diagnostics.

Figure 1. Network of all handheld devices (diagnostic and others) providing healthcare administration and healthcare information management systems-related services to doctors and patients. A single handheld diagnostic device with all the diagnostic functionality enables the patient to get desired healthcare services.

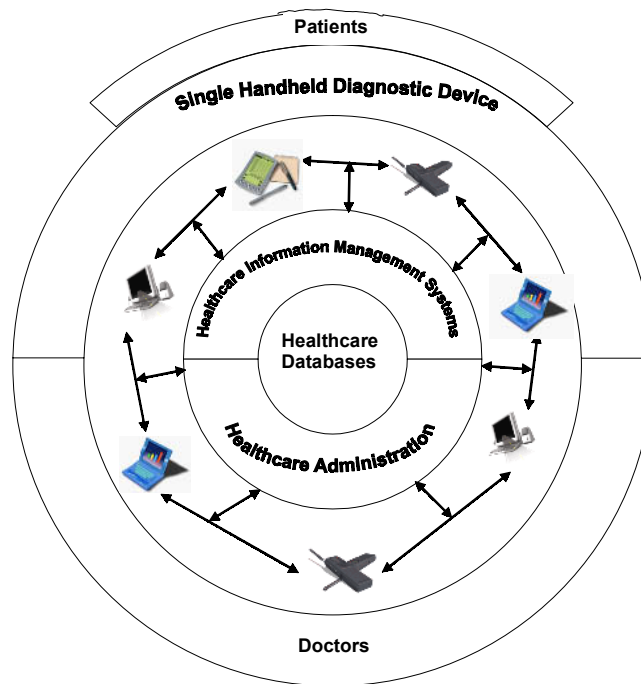
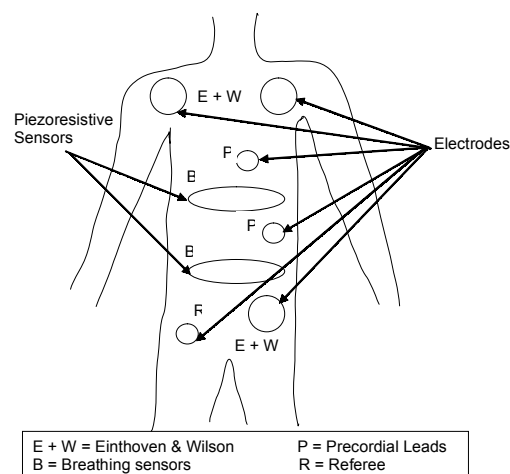


Figure 2. A representative diagram of WEALTHY



NEXT STEPS

We are at the crossroads of mobile technologies and healthcare diagnostics today. While both of them converge here coming from two different directions, the convergence itself is expected to advance in two different directions simultaneously. These two directions are as follows:

1. **Integration of all the handheld diagnostic devices or their functionality** into a single device that is connected to the wired and wireless networks. Eventually, they would be interconnected with healthcare networks and databases (generic as well as medical records). This model would look somewhat similar to that in Figure 1.

In fact, European researchers have just completed something very similar—the Wearable Health Care System project, otherwise known as WEALTHY (2005). In addition to being fashionable and slimming, the body suit will measure all manner of bodily processes, including respiration, core and surface skin temperature, position, and movement. It will also transmit all this data over the mobile phone network thanks to a miniaturized GPRS transmitter hidden in the suit. The transmitter can also alert emergency services if the wearer stops responding or shows alarming vital signs. WEALTHY is depicted in Figure 2.

2. **Networking of individual handheld diagnostic devices** amongst themselves and to wired and wireless networks. Eventually, they would also be interconnected with healthcare networks and databases (generic as well as medical records). This model would look somewhat similar to that shown in Figure 3.

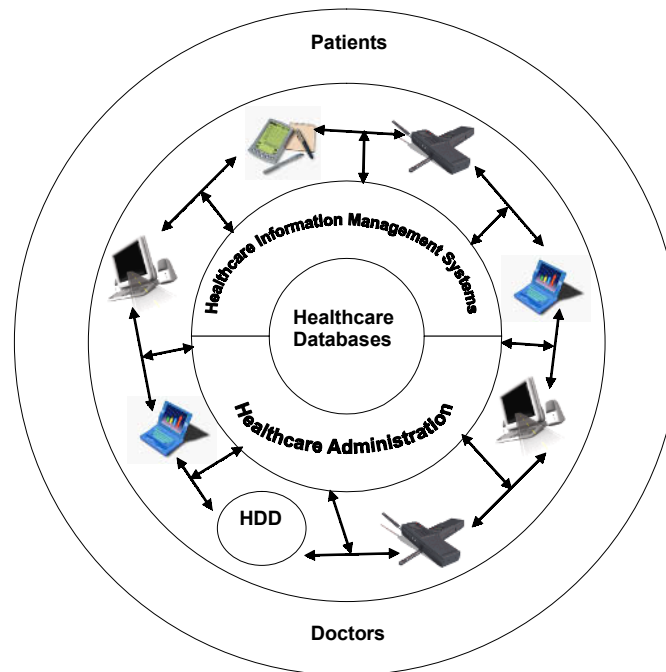
MOBILE TECHNOLOGIES IN HEALTHCARE ADMINISTRATION

Healthcare administration consists of all the activities undertaken to provide healthcare services to patients. Broadly, these activities are:

- **Consultation**
 - Patient visits the doctor
 - Patient provides details of their ailment to the doctor
 - Doctor carries out physical examination for symptoms
 - Doctor may suggest further diagnostic tests—the probability of this is much higher today with the advancement of medical science
- **Diagnosis**
 - Patient visits various laboratories for diagnostic tests
 - Patient visits the laboratories again to collect the diagnostic test reports
 - Patient visits the doctor again with the diagnostic test reports
- **Treatment**
 - Doctor prescribes medication to the patient or recommends a surgery
 - Patient visits the pharmacy to buy the medication
 - If recommended, patient registers for the surgery at a hospital
- **Payment**
 - Patient makes payment to the doctor
 - Patient makes payment to the pharmacy
 - Patient approaches the insurance company or concerned agency for reimbursement of medical expenses.
 - Alternatively, the medical agencies (like the physician) approach the insurance company for reimbursement

What we need to explore now is how many of these steps can be performed effectively

Figure 3. Network of all handheld devices (diagnostic and others) providing healthcare administration and healthcare information management systems-related services to doctors and patients. HDD = a network of handheld diagnostic devices that is integrated with wired and wireless networks. This entire network is accessible to doctors and patients.



using mobile devices like cell phones or PDAs. I would not be overstating if I mentioned that most of the activities related to healthcare administration can be performed using various handheld devices. Table 1 shows how.

ADDITIONAL THOUGHTS

In addition to the normal healthcare administration process, there are many more things that can be achieved using handheld technologies. Some of them are:

- *Embedding important patient data on the smart cards* of mobile phones or PDAs. In cases of emergency like accidents, war, and so forth, quite often the patient is unable to provide any basic information about himself that is important for the doctor to initiate treatment. All such information, starting from the blood group of the patient, his address, contact details of his immediate relatives, his allergies, a brief medical history consisting of all major healthcare-related events in the patient's life, and so on can be embedded on the smart card so that it is accessible at the click of a button. This can enable much faster treatment of the patient.
- *GPS enabling of handheld devices.* This also helps in tracking a patient during emergencies. On the diagnostics front, there are already devices that can keep track of a patient's health on certain key

Application of Mobile Technologies in Healthcare Diagnostics and Administration

Table 1.

Activity	How to perform it using handheld device/s – there could be multiple options
Consultation • Patient visits the doctor • Patient provides details of their ailment to the doctor • Doctor carries out physical examination for symptoms • Doctor may suggest further diagnostic tests	• A personal visit by the patient to the doctor may not be required in many cases. Patient can talk to the doctor using cell phones and provide details of their ailment. • If needed, patient can also have a video conference with the doctor using cell phones with cameras & multi-media capabilities or smart phones – this would enable the doctor to observe the physical symptoms of the patient's ailment. • If a meeting with the doctor is absolutely necessary, a patient can access the doctor's calendar through their cell phone or PDA and take an appointment. Similarly doctors too can maintain and access their calendar through their cell phones or PDAs. • Having understood the ailment as well as symptoms, doctors can send a recommendation or prescription for diagnostics to the patient. Such a prescription can be created quickly using a pre-loaded form on the handheld device and then sent to the handheld device of patient as well as to the relevant laboratories.
Diagnosis • Patient visits various laboratories for diagnostic tests • Patient visits the laboratories again to collect the diagnostic test reports • Patient visits the doctor again with the diagnostic test reports	• As discussed at length in the previous section, most of the diagnostics can now be done through handheld devices. A patient in this case would have 3 choices – 1. If he is suffering from chronic ailment, he may have his own diagnostic device/s that he can use for the test and then send the result to the doctor, through a mobile device. 2. Patient can visit a laboratory to get the tests done. The laboratory, in turn, can flash the results to the handhelds of the doctor as well as the patient as soon as they are ready, saving the patient of a second visit. 3. With so many small handheld devices available, there is a good possibility of a doctor having them. Thus, patient can just pay one visit to the doctor and complete consultation, physical examination of symptoms as well as diagnostics – all at a time. • In all the cases, the patient and doctor can achieve a lot with more convenience using handheld devices.
Treatment • The doctor prescribes medication to the patient or recommends a surgery • Patient visits the pharmacy to buy the medication • If recommended, patient registers for the surgery at a hospital	• After studying diagnostic reports, doctor can prescribe medication and therapies to the patient. This too can be done using their respective handheld devices. • A doctor can also take help from a lot of references (medical guides, reference, encyclopedia, etc.) preloaded on the handheld or accessible on the net through the handheld. • The prescription of medication and therapies duly signed by doctor can be transmitted to the patient's handheld device. • Patient can visit any pharmacy with the preloaded prescription from the doctor on the handheld device and buy the medicines by providing the prescription to the pharmacy. • If a surgery is recommended, the patient can take an appointment with the concerned doctor and surgery by accessing their calendar through the handheld. • In the surgery, if needed, the doctor/s can get medical advice / expertise from various resources on the net using their handheld devices or even having real-time interaction with experts around the world.
Payment • Patient makes payment to the doctor • Patient makes payment to the pharmacy • Patient approaches the insurance company or concerned agency for reimbursement of medical expenses. • Alternatively, the medical agencies (like the physician) approach the insurance company for reimbursement.	• Patient can pay the doctor / pharmacy using preloaded cash or credit/debit cards on the handheld device. • Patient can similarly receive all the invoices on his handheld and forward them to the insurance company along with a filled in prescribed form for reimbursement. • In case, the doctor has to approach the insurance agencies, they too can forward the relevant forms and documents in digital format from their handheld devices.

parameters and notify the doctor if a negative trend is building on any of them. An integration of these devices with GPS-enabled handheld devices would add a lot more value to provide proactive healthcare to patients.

Thus, a large part of healthcare administration can actually be carried out using handheld devices. One of the primary enablers for this is that the latest models of handheld devices are turning out to be as capable as desktops and laptops of not so long ago—virtually any task that one could complete on a desktop or laptop 3-5 years ago can now be accomplished through a handheld device. In fact, almost all the technologies are already available. All that is required is their integration or convergence.

MOBILE TECHNOLOGIES IN HEALTHCARE INFORMATION SYSTEMS

Having discussed how mobile technologies facilitate healthcare administration, which is essentially interactions between a healthcare service provider—a doctor or a clinic—and a patient for providing healthcare services to the patient, let us now see how these technologies can be useful in healthcare information systems.

Healthcare administration results in the generation of a lot of data that needs to be maintained and managed for future reference. Such data includes, but is not limited to, EMRs (electronic medical records), diagnostic reports, and patients' medical history. Healthcare service providers like doctors, clinics, and hospitals also need to maintain a lot of additional information on schedules of doctors (for appointments) and resources (like operation theatre, etc.), infor-

mation related to finances and inventories, references on the latest developments in the healthcare industry through subscription to various databases, and so on. Such information needs to be readily accessible, anytime, anywhere. Such information also needs to be available for research studies, study of trends and patterns, and so forth.

Healthcare information systems consist of creation, maintenance, provisioning, and retirement of all such information related to healthcare. Typically this would mean maintaining a healthcare information management system (HIMS) in a clinic or hospital. While this is true in today's world, Vanjara (2005) has suggested a geography-time-person agnostic solution. The solution consists of creating networked virtual healthcare communities that have access to relevant information anytime, anywhere.

In both the scenarios, information can be created, maintained, accessed, and retired through handheld devices. Natural interfaces have further removed one of the major constraints in using handheld devices through the use of keyboards or mice. In addition, there are thousands of software applications and utilities, many of them available as freeware on various handheld platforms that enable mobile healthcare information systems.

Healthcare information can be broadly classified into two types:

1. **Medical Information:** This consists of all information pertaining to a patient's health and includes EMRs (electronic medical records), test reports with images as applicable, history of ailments, diagnostics and treatment, patterns and profile of a patient's health, and so on.
2. **Non-Medical Information:** This consists mostly of all commercial information like receivables and payables, accounts,

inventory, and so on. In case of hospitals, such information would be much vaster, including data around various administrative and support functions like laundry, kitchen, and so on.

Currently, only mid-sized to large clinics or hospitals maintain comprehensive real-time healthcare information management systems. This scenario has two constraints: (1) the smaller entities do not have any system for HIMS, and therefore data/information related to patients' interactions with them is never maintained; and (2) even the HIMS in mid-sized or large clinics and hospitals is unable to capture and maintain medical information of patients from different sources. Imagine test reports from four different places being compiled in a single central database of a hospital! Sounds quite difficult.

Mobile devices enable HIMS in both cases. In case of smaller entities, they do not have to make large investments in infrastructure to have a comprehensive HIMS. They can actually subscribe to such infrastructure and Web-based applications, and then access them to maintain their information using handheld devices. The HIMS of mid-sized and large clinics and hospitals can be easily fed with all data from multiple sources almost on a real-time basis using handheld devices. Even voice-based interactions/transactions can be captured using the voice recognition feature of mobile devices. The primary idea behind handheld devices is to capture as much data as possible in digital format right from the point of origin. Manipulation and transfer of data after that becomes quite easy and fast.

Though maintaining all types of information—medical or non-medical—is possible using handheld devices, medical information is a primary candidate because it addresses more issues around mobility than non-medical information, which is to a large extent organization or location specific.

ISSUES AND CHALLENGES

Having read and heard so much about the potential and capabilities of mobile/handheld devices, the first obvious question that comes to mind is, why then are they not used so commonly in these areas of healthcare? Growth of mobile technology advancements in the entertainment and consumer sectors has been much higher than in business. Given the fact that it is business that actually contributes to the economy, and in case of healthcare, also to the fitness and well-being of people, this is quite unfortunate.

However, there are some reasons for these as well—in the form of issues and challenges of mobile technologies:

- **Integration of Devices, Applications, and Platforms:** As of now, most of the devices, applications, and platforms work independently. Even mobile phones of two different manufacturers at times cannot communicate with each other. At most, they exchange some data via IR or Internet; however, there is absence of tight integration in terms of features, functionality, and so on. A robust, standard-based integration of multiple devices, applications, and platforms would go a long way in facilitating wider use of mobile devices in the business of healthcare.
- **Investments and Infrastructure:** A lot of investments need to be made in the non-telecom space, including creation of application infrastructure to popularize the use of mobile devices. This would also entail creation of a lot of enterprise-scale application software that can expose selected services to multiple clients. To work at a global scale, at each level or tier the data will have to be transparent to devices and technologies. This again will require investment in technologies and standards.

- **Common Standards:** Handheld devices, software applications and platforms, and the healthcare industry also need to have common standards of communication amongst themselves, like the MICS—the Medical Implant Communications Service, HL7, and so on. Such standards would enable multiple devices to talk to multiple software/services on different platforms.
- **Privacy and Security:** While handheld device space is not as rampant with these issues today as is the personal computer space, it will not be long before it catches up. Some cases of viruses on mobile phones and PDAs or data pilferage or snooping have already been reported. Privacy and security become of paramount importance in the healthcare space due to two primary reasons: (1) we are constantly dealing with individuals' personal information, and (2) the impact of security breach can cost a human life. Imagine a software virus that keeps on manipulating prescriptions or for that matter diagnostic reports while spreading itself.
- **Laws:** Laws related to the conduct of various industries like healthcare are also widely different in various countries. These too need to be aligned at a broad level for a universal solution to work. While this is very difficult and far fetched, if all the countries in the world can sign charters and conventions on pollution control, IPRs, and many such issues, creation of a common legal framework at a higher level for the benefit of mankind—for instance in the area of healthcare—can certainly be made possible in the long term.
- **Language:** A global solution also has to address the need of multiplicity of languages. This, though, is the least of problems, as quite a few solutions—Microsoft Windows to small accounting packages

like QuickBooks and MYOB—have already addressed this issue. Unicode-based solutions can also be considered to address this issue (<http://www.unicode.org>)

- **Need for Social Interaction:** Most of the technological advancements have resulted into lesser human-to-human interactions. This is creating a large gap in the social needs of people. Quite often, at a sub-conscious level, people express resistance to new technologies due to their adverse impact on the fulfillment of their social need.

CONCLUSION

This chapter reviews the healthcare and mobile technologies landscape as it exists today. It then suggests integration of both these technologies—already happening to some extent—for provisioning of superior healthcare services in a flexible manner from anywhere. It discusses how various mobile healthcare gadgets and communication devices can be networked and connected to healthcare administration and information systems. Having proposed a healthcare solution that is a few quantum jumps from where it exists today, the chapter also discusses some issues and challenges in making this happen. It also makes a few suggestions on how some of these issues and challenges can be taken care of.

REFERENCES

- Apple Computer, Inc. (2005). *iPod. Which iPod are you?* Retrieved July 27, 2005, from <http://www.apple.com/ipod/>
- Cellular Network. (2005). Retrieved August 3, 2005, from http://www.mywiseowl.com/articles/Cellular_network

- Computerworld. (2005). *Combo Wi-Fi cell phone*. Retrieved August 3, 2005, from <http://www.computerworld.com/mobiletopics/mobile/story/0,10801,93501,00.html?nas=AM-93501>
- Dixon, P. (2002). *Everything you wanted to know about the future of health care industry. Keynote speech*. Retrieved March 10, 2005, from <http://www.globalchange.com/futuremedicine.htm>
- EIT.IHK. (2005). *Man machine interface*. Retrieved March 12, 2005, from <http://www.eit.ihk-edu.dk/subjects/mmi/index.php?e=0>
- Free Medical Journals. (n.d.). Retrieved from <http://www.freemedicaljournals.com/html/special.htm>
- Global Change. (2005). *The future of technology. Recent news of a new kind of world*. Retrieved April 30, 2005, from <http://www.globalchange.com/futuretechnology.htm>
- Google. (2005). *Google directory-health > news and media > publications > journals*. Retrieved July 27, 2005, from http://www.google.com/Top/Health/News_and_Media/Publications/Journals/
- Handango. (2005). *Handheld software for healthcare*. Retrieved April 10, 2005, from http://www.handango.com/SoftwareCategory.jsp?optionId=1_2_2&jid=5232151DE62C1F9XXX94X1785XD761AX&special=&platformId=2&bySection=1&siteId=1&txtSearch=medical&sectionId=3314&topSectionId=3314&catalog=30&title=Medical
- Health-Cares. (2005). *What causes a common cold?* Retrieved July 27, 2005, from <http://respiratory-lung.health-cares.net/cold-causes.php>
- HealthDay. (January 21). *High-tech revolutionizes home health care*. Retrieved March 13, 2005, from <http://www.azcentral.com/health/wellness/articles/0121hightechhealth-ON.html>
- Howard, V. (2004). Nanotechnology and medicine. *Ecologist Asia*, 12(1). Retrieved July 27, 2005, from <http://www.sanctuaryasia.com/features/detailfeaturescategory.php?id=666&catid=45>
- HP iPAQ 6315. (2005). *Product description*. Retrieved July 27, 2005, from <http://h10010.www1.hp.com/wwpc/us/en/sm/WF06b/215348-64929-215381-314903-f60-430120-430121-405360.html>
- Light Blue Optics (2004). *About us*. Retrieved July 27, 2005, from <http://www.lightblueoptics.com/index.html>
- Majeed, M. (2005, March 10). *Recent advances in medicine (speech at Trivandrum Medical College)*. Retrieved April 15, 2005, from <http://www.drmajeed.com/speeches.htm#Recent>
- Majumdar, A. (2004). *Ayurveda: The ancient Indian science of healing*. Delhi: Macmillan.
- Medgadget. (2005a). *Amplichip*. Retrieved March 12, 2005, from http://www.medgadget.com/archives/2005/01/amplichip_1.html
- Medgadget. (2005b). *Fantastic voyage: Smart Pill to expand testing*. Retrieved March 12, 2005, from http://www.medgadget.com/archives/2005/04/fantastic_voyag.html
- Microsoft. (2005). *Product description of Windows Mobile and MS Office Mobile*. Retrieved July 27, 2005, from <http://www.microsoft.com/windowsmobile/default.mspx> and <http://office.microsoft.com/en-us/assistance/HA011017991033.aspx>
- Microsoft Streets & Trips 2005 GPS. (2005). Retrieved April 30, 2005, from <http://www.tigerdirect.com/applications/searchtools/>

item-details.asp?EdpNo=1016599&Sku=M17-5667&SRCCODE=MSN&CMP=OTC-MSN

PDA Buzz. (n.d.). Retrieved from <http://www.pdabuzz.com/>

PDA Handyman. (n.d.). Retrieved from <http://www.pdahandyman.com/>

PDA Live. (n.d.). Retrieved from <http://www.pdalive.com/>

PMED (2005). *Patient monitors, defibrillators, telemetry, stress test, vascular*. Retrieved March 7, 2005, from <http://www.pemed.com/pmonitor/pmon.htm>

Ratib, O. (2004). iPod helps radiologists manage medical images. *RSNA News*, (December). Retrieved March 7, 2005, from <http://www.rsna.org/publications/rsnanews/dec04/ipod-1.html>

Shaheen, M. (2003). *How does WiFi work?* Retrieved August 3, 2005, from <http://mason.gmu.edu/~mshaheen/page1.html>

Unicode. (2005). *Unicode home page*. Retrieved March 10, 2005, from <http://www.unicode.org/>

Vanjara, K. (2005). The future of supply chain management: Shifting from logistics driven to a customer driven model. In Y. Lan & B. Unhelkar (Eds.), *Global integrated supply chain systems*. Hershey, PA: Idea Group Publishing (submitted for publication).

WEALTHY. (2005). *Description*. Retrieved March 12, 2005, from <http://engadget.com/entry/1234000690040296/#comments>

Section III

Technical

Chapter X

Energy–Efficient Cache Invalidation in Wireless Mobile Environment

R. C. Joshi

Indian Institute of Technology Roorkee, India

Manoj Misra

Indian Institute of Technology Roorkee, India

Narottam Chand

Indian Institute of Technology Roorkee, India

ABSTRACT

Caching at the mobile client is a potential technique that can reduce the number of uplink requests, lighten the server load, shorten the query latency and increase the data availability. A cache invalidation strategy ensures that any data item cached at a mobile client has same value as on the origin server. Traditional cache invalidation strategies make use of periodic broadcasting of invalidation reports (IRs) by the server. The IR approach suffers from long query latency, larger tuning time and poor utilization of bandwidth. Using updated invalidation report (UIR) method that replaces a small fraction of the recent updates, the query latency can be reduced. To improve upon the IR and UIR based strategies, this chapter presents a synchronous stateful cache maintenance technique called Update Report (UR). The proposed strategy outperforms the IR and UIR strategies by reducing the query latency, minimizing the disconnection overheads, optimizing the use of wireless channel and conserving the client energy.

INTRODUCTION

The tremendous growth in mobile hardware technology and wireless communication has

increased the number of clients that access data remotely. Efficient data access in mobile computing is a field of increasing importance for a wide range of mobile applications. Users

of mobile devices wish to access dynamic data, such as stock quotes, news items, current traffic conditions, weather reports, e-mail, and video clips via wireless networks. However, limited battery power of mobile client and scarce wireless bandwidth hinder the full realization of ubiquitous data access in mobile computing. Caching at the mobile client can relieve bandwidth constraints imposed on wireless mobile computing. Copies of remote data can be kept in the local memory of the mobile client to substantially reduce user requests for retrieval of data from the origin server. This not only reduces the uplink and downlink bandwidth consumption, but also the average query latency. Caching frequently accessed data by a mobile client can also save its power used to retrieve the repeatedly requested data.

Cache invalidation strategy is used to ensure that the data items cached at a mobile client are consistent with those stored on the server. Depending on whether or not the server maintains the state of the mobile client's cache, the invalidation strategies are divided into two categories: the *stateful* server approach and the *stateless* server approach (Barbara & Imielinski, 1994; Tan, Cai, & Ooi, 2001). Barbara and Imielinski (1994) provide a solution where the server periodically broadcasts an invalidation report (IR) in which the changed data items are indicated. Rather than querying the server directly regarding the validation of cached copies, the clients can listen to these IRs over the wireless channel and use them to validate their local cache. The IR-based invalidation may be of two types: *synchronous* and *asynchronous*. In the synchronous method, the invalidation reports are broadcast periodically, whereas in the asynchronous method, the server broadcasts the reports only when some data changes. Because of the nature of periodic broadcast, synchronous methods provide a bound on the waiting time of the next report, whereas in an asynchronous invalidation re-

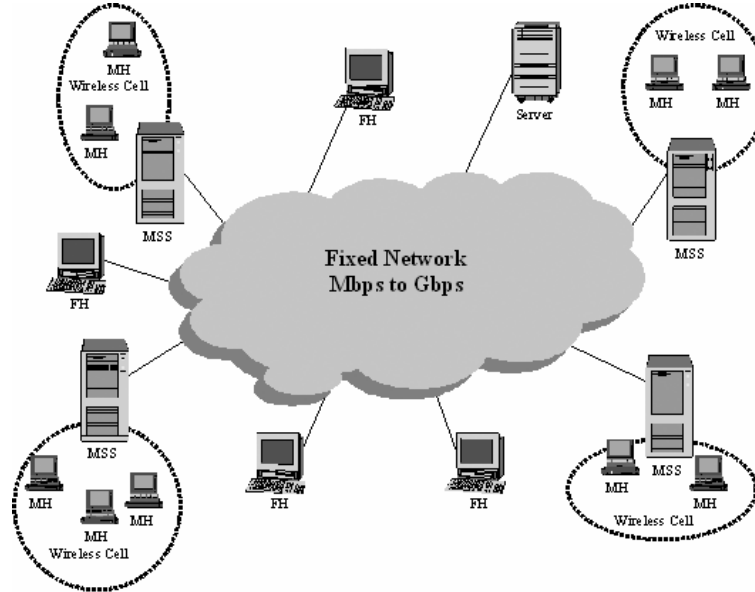
port, there is no guarantee on how long the client must wait.

Clients use IRs to keep their cache consistent by discarding any obsolete data. If a query cannot be served locally—that is, a cache miss—the client issues an uplink query request for the data items. The IR-based solution is attractive because of its scalability, as the size of IR is independent of the number of clients. It is also energy efficient, as clients can exploit the periodicity of server broadcast to save power, in that mobile devices can operate in doze mode most of the time and only activate during broadcast. However, the solution suffers from the problem of long query latency since a client must listen to the next IR before answering a query. The problem has been tackled with the addition of *updated invalidation report (UIR)* by broadcasting a number of smaller reports (UIRs) between successive IRs (Cao, 2001, 2002a, 2002b, 2003). Each UIR contains information about most recently updated data items since the last IR. In case of cache hit, there is no need to wait for the next IR and hence the query latency is reduced. However, if there is a cache miss, the client still needs to wait for the data to be delivered. Thus, due to cache miss, the UIR strategy has the same query latency as IR strategy.

In IR strategy, if the disconnection time of a client is longer than a fixed period, the client should discard its entire cache even if some of the cached data may still be valid. This issue is addressed in Cao (2002a, 2002b), and Jing, Elmagarmid, Helal, and Alonso (1997). Chand, Joshi, and Misra (2005) have demonstrated more efficient handling of arbitrarily long client disconnection.

To overcome the limitations of existing cache invalidation strategies, we present a synchronous stateful caching strategy where cache consistency is maintained by periodically broadcasting *update reports (URs)* and *request reports (RRs)*. The central design of our strat-

Figure 1. UR caching model



egy includes reducing the query latency, improving the cache hit ratio, minimizing the client disconnection overheads, utilizing the wireless channel better, and conserving the client energy. The track of cached items for each client is maintained at the home mobile support station in the form of *cache state information (CSI)*. Use of CSI reduces the size of IR by filtering out non-cached items and handles long disconnection. In various IR-based strategies (Kahol, Khurana, Gupta, & Srimani, 2001; Jing et al., 1997; Barbara & Imielinski, 1994; Chuang & Hsu, 2004), even though many clients cache the same updated data item, all of them have to query the server and get the data separately from the server. It wastes a large amount of wireless bandwidth and client battery energy. To minimize uplink requests and downlink broadcasts, we use a broadcast strategy, called update report (UR) (Chand et al., 2005), where all the recently updated/requested items are broadcast immediately after the invalidation report (IR). To further reduce query latency, the

strategy uses *request reports (RRs)*, where all the recently requested items are broadcast after the UIR. Selective tuning is used to conserve the client energy.

THE PROPOSED CACHE INVALIDATION STRATEGY

In this section, we present our UR-based synchronous stateful caching strategy.

UR Caching Model

As shown in Figure 1, the model consists of two distinct sets of entities: *Mobile Hosts (MHs)* and *Fixed Hosts (FHs)*. Some of the fixed hosts, called *Mobile Support Stations (MSSs)*, are augmented with a wireless interface in order to communicate with the mobile hosts, which are located within a radio coverage called a cell. Each cell is associated with an id for identification purpose. MSSs are also known

as *Base Stations (BSs)*. An MSS acts like a gateway between a fixed network and a wireless network. An MH communicates with a fixed host/server via an MSS over a wireless communication link. The communication is asymmetric (i.e., the uplink bandwidth is much less than that of downlink). The MSSs communicate among themselves over a wired channel and the communication is transparent to a client. A fixed network has a large bandwidth (order of Mbps or Gbps), while the bandwidth of the wireless channel is low (19.2 Kbps-10 Mbps). An MH can move within a cell or between cells while retaining its network connection. When an MH moves from one cell to another (called handoff), its wireless connection is switched to the new cell. An MH either connects to an MSS through a wireless link or disconnects from the MSS by operating in a 'power save' mode (Kahol et al., 2001).

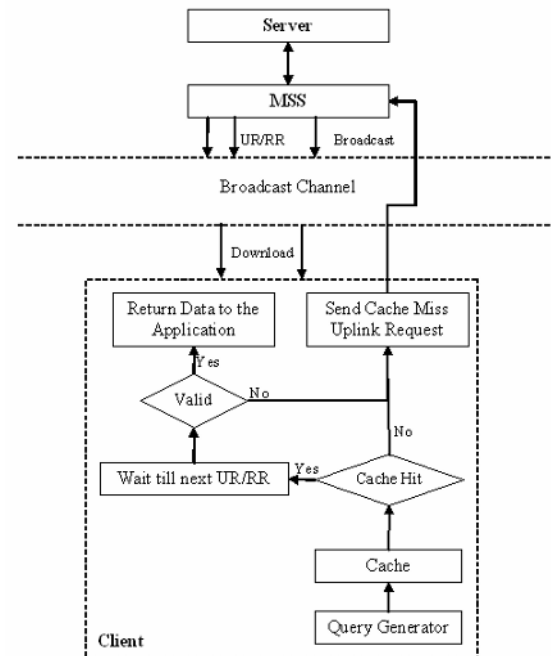
The database D is a collection of N data items with ids: $d_1, d_2, \dots, d_N$. A data item is the basic unit for update and query. For each data item d_i , two timestamps t_i and t_i^r are maintained: t_i is the most recent *timestamp* when d_i got updated at the server and t_i^r , called *latest request time*, represents the most recent time when d_i was last requested by any client. MHs only issue simple requests to read the most recent copy of a data item. In order to serve a request sent from a client, the MSS needs to communicate with the database server to retrieve the data items. Caching techniques may also be applied at MSS. Since the communication between the database server and MSSs are through wired link, we assume traditional techniques can be used to maintain cache consistency. Since the communication between the MSS and the database server is transparent to the clients, from the client point of view, the MSS is the same as the database server.

Frequently accessed data items are cached on the client side. We assume that the cache at

the mobile client is a nonvolatile memory such as a hard disk so that after a long disconnection, the contents of the cache can still be retrieved. When caching is used, data consistency issues must be addressed. We assume the *latest value* consistency model (Cao, 2002a, 2002b), which is widely used in dissemination-based information systems.

To ensure cache consistency, the server broadcasts UR every L seconds and it also broadcasts $(m-1)$ RRs between two URs. Every active client listens to the report (UR/RR) and invalidates its cache accordingly. To answer a query, the client listens to the IR/UIR part of the next report (UR/RR) and decides its cache validity. If there is a valid cached copy of the requested item, the client returns the item immediately. Otherwise, it sends a query request to the server through the uplink. The simulation architecture of the proposed model is shown in Figure 2.

Figure 2. Simulation architecture



To keep the information about cached items for each MH, a cache state information (CSI) is maintained at the home MSS. The CSI is the list of cached data item ids by the host. For each item d_i , a cache count n_i is also maintained at the home MSS. Thus, n_i denotes the number of clients who have cached the item d_i in that particular cell. When a client sends a data request, the MSS updates the relevant counters and the corresponding CSI, and forwards the request to the server.

In order to save energy, an MH may power off most of the times and only turns on during the report broadcast time. Moreover, an MH may be in the power save mode for a long time and it may miss some reports.

The following assumptions are made:

- Database D is a collection of N data items. An item is identified by a unique id d_i ($1 \leq i \leq N$). D_i denotes the actual data of an item with id d_i . Each item has the same size S_{data} (in bits).
- Each cell has a single MSS such that cell A is managed by MSS_A . Each MSS broadcasts UR every L seconds and RR every L/m seconds.
- A unique host identifier is assigned to each MH in the system. The system has a total of M hosts, and MH_i ($1 \leq i \leq M$) is a host identifier. Each mobile host moves freely. We use the terms host and client interchangeably.
- Each MH has a cache space for C data items.
- CSI stored in the local disk of home MSS maintains the state information for a host. An MH informs its MSS before it stores any data item in its local cache and the MSS updates the CSI accordingly.
- The server is reliable—that is, it handles the failure with some fault tolerance techniques.

Cache State Information to Reduce the Report Size

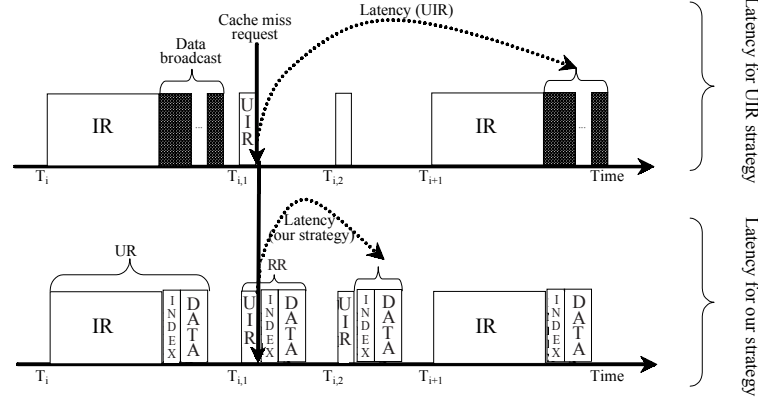
In a stateless strategy when an item updates at the server, its id is broadcast as part of IR irrespective of whether the item has been cached or not. Including an item as part of IR, which has not been cached by any client, makes poor utilization of the available wireless bandwidth. It also increases the client energy consumption since users have to listen to the broadcast channel for a longer duration to download the report. To filter out from a report all those recently updated items that are not cached by any client, we have used a stateful approach in our strategy.

To keep the information about cached items, for each MH a *cache state information (CSI)* is maintained at the home MSS. Consider a cell with H hosts (MH_i , $1 \leq i \leq H$), at any given time. For any j , CSI_j for MH_j , as maintained on its home MSS, keeps track of what data has been locally cached at MH_j . In general, if $d_k \in CSI_j$, then the client MH_j has cached the item d_k . When an item updates, it will be added as part of a report (UR/RR) at the server. The MSS, upon receiving the report from the server, removes all those items from the IR that are not cached by any client (i.e., with cache count 0) and thus broadcasts reduced report in its cell. When a client moves to a new cell, the copy of its CSI is replicated at the new MSS.

Reducing the Query Latency

In UIR-based caching strategy, the server aggregates data requests from all its clients over the whole invalidation interval (L seconds) and broadcasts the requested data after each IR. This aggregation of requests tremendously reduces the number of data broadcasts and thus makes efficient utilization of the downlink channel. The reduction in the number of broadcasts

Figure 3. Reducing the query latency



is at the expense of increased query latency, since a client has to wait longer to download the requested data item. In UIR scheme, the requested data are scheduled for broadcast after the next IR, thus due to cache miss, the expected query latency is $L/2$ seconds. To reduce the query latency due to cache miss, the UR strategy broadcasts the recently requested data items after the next report (IR/UIR), whichever arrives earlier, such that the expected query latency is $L/(2*m)$ seconds instead of $L/2$ seconds. In general, $T_{i,k}$ represents the time of k^{th} RR after the i^{th} UR. When a client receives a cache miss request between $T_{i,1}$ and $T_{i,2}$, it cannot answer the query until T_{i+1} in the UIR approach, but it can answer the query at $T_{i,2}$ in UR approach (see Figure 3). UIR, followed by the broadcast of recently requested data, constitutes *request report* (RR). At interval time $T_{i,k}$, $RR_{i,k}$ is constructed as follows:

UIR _{i,k}	RR_INDEX _{i,k}	RR_DATA _{i,k}

$$\begin{aligned} \text{UIR}_{i,k} &= \{d_x | (d_x \in D) \wedge (T_{i,0} < t_x \leq T_{i,k}) \wedge (n_x > 0)\} \quad (0 < k < m) \\ \text{RR\_INDEX}_{i,k} &= \{d_x | (T_{i,k-1} < t_x \leq T_{i,k})\} \\ \text{RR\_DATA}_{i,k} &= \{D_x | d_x \in \text{RR\_INDEX}_{i,k}\} \end{aligned}$$

This distribution of query replies also reduces the impact of data broadcast on other downlink traffic. To make the selective tuning possible for the clients, the server broadcasts the index information RR_INDEX before the broadcast of actual data. Since the query replies are distributed, the size of the index in our strategy is much smaller than in the UIR strategy.

Improving Wireless Channel Utilization

To reduce the number of uplink requests and downlink broadcasts, we introduce the concept of update report (UR) (Chand et al., 2004). Update reports (URs) are broadcast synchronously with period L .

At interval T_i , the structure of UR_i is as follows:

IR _i	UR_INDEX _i	UR_DATA _i

$$\begin{aligned} \text{IR}_i &= \{(d_x, t_x) | (d_x \in D) \wedge (n_x > 0) \wedge (T_i - w*L < t_x \leq T_i)\} \\ \text{UR\_INDEX}_i &= \{d_x | ((T_{i-1} < t_x \leq T_i) \wedge (n_x > 0)) \vee (T_{i-1,m-1} < t_x \leq T_i)\} \\ \text{UR\_DATA}_i &= \{D_x | d_x \in \text{UR\_INDEX}_i\} \end{aligned}$$

UR_INDEX_i defines the order in which data appears in UR_DATA_i . Within UR_INDEX and RR_INDEX , the items are arranged in non-decreasing order of their cache count. This ordering of broadcast items further reduces the query latency (Chand et al., 2004).

IR contains the update history of past w broadcast intervals, whereas UR_DATA contains the actual data value for the items that have been updated during previous UR interval and the items that have been requested during the last RR interval. In our strategy the contents of URs broadcast in different cells depend upon the cache state of the clients lying within a cell, and hence the broadcast URs may be inhomogeneous. In most IR-based algorithms (Kahol et al., 2001; Jing et al., 1997; Barbara & Imielinski, 1994), updating a data item that has been cached may generate many uplink requests and downlink broadcasts, and thus make poor utilization of available wireless bandwidth. This is due to the reason that when an item is updated and IR is broadcast, each client who has cached that item will generate an uplink request for the item and the server responds to each request by broadcasting the item. For example, for an item with id d_x which is cached by n_x clients, there will be n_x uplink requests and n_x downlink broadcasts due to update. We address the problem by asking the server to broadcast all the data items that have been recently requested or updated and are cached by one or more clients. If a client observes that the server is broadcasting an item which is an invalid entry in its local cache, it will download the item. Otherwise, the client may have to send another request to the server, and the server will have to broadcast the data again in the future. So in comparison to n_x uplink requests and downlink broadcasts for an updated item, our strategy makes only single broadcast without any uplink request.

Due to data update at the server, UR strategy has same number of uplink requests and downlink broadcasts as in UIR strategy. Also, during one RR interval, due to cache miss an item may have been requested by many clients, but our scheme broadcasts the item only once. In comparison to UIR strategy, our strategy further improves the wireless channel utilization by using *delayed uplink (DU)* technique as follows.

To improve the efficiency of uplink channel due to cache miss, when an item d_x is requested by a client at time T_{ij} ($1 \leq j \leq m-1$), in UR strategy the client would download RR_INDEX_{ij} to see if the server has planned the broadcast of item d_x . If $d_x \in RR\_INDEX_{ij}$, the client would download D_x and the item id d_x is piggybacked when a new request is sent to the server; otherwise the client sends an uplink request to the server for d_x . This saving in uplink request also reduces query latency as the client receives the item sooner. Thus, reducing the number of uplink requests and downlink broadcasts due to recent updates or cache misses, UR strategy heavily saves on wireless bandwidth.

Synchronous Broadcasting to Conserve Client Energy

In asynchronous invalidation methodology, there is no guarantee on how long the client must wait for the next report, and hence the clients are in doze mode and may lose some of the reports, thus compromising the cache consistency or further increasing the query latency. By broadcasting UR and RR periodically, we use a synchronous approach where clients may wake up during the UR/RR broadcast time and selectively tune in to the channel to save power. After broadcasting IR/UIR, the server broadcasts UR_INDEX/RR_INDEX followed by the broadcast of actual data $UR\_DATA/$

RR_DATA. Every client listens to the report (IR/UIR) if not disconnected. At the end of report, the client downloads index and locates the interesting item that will come, and listens to the channel at that time to download the data. This strategy saves power since the client selectively tunes to the channel and can stay in doze mode most of time.

Handling Client Disconnection

Since a UR broadcasts information about the items that have been updated during past $w*L$ time, our strategy handles the disconnection of clients less than $w*L$ without any additional overhead. When a client reconnects after a disconnection time longer than $w*L$, it sends an uplink request with the last received UR time stamp T_1 (before disconnection) to the home MSS. On receipt of the request, the MSS constructs a binary vector *DIV* called *disconnection information vector*. *DIV* is of size C bits and contains the validity information about the cached items by the client. For a client MH_i , the MSS constructs DIV_i as follows:

1. Scan the CSI_i for the list of cached items. If $d_j \in CSI_i$, MH_i has cached the item d_j otherwise not ($1 \leq j \leq N$).
2. For an item d_j which is cached by client MH_i , compare its last update timestamp (t_j) with T_1 . If $t_j > T_1$, the item d_j has been updated since MH_i received the last UR before disconnection. In case t_j satisfies the above condition (i.e., $t_j > T_1$), then set $DIV_i[k] = 1$, where MH_i has stored item d_j at k^{th} cache location ($1 \leq k \leq C$). If $t_j \leq T_1$, then set $DIV_i[k] = 0$.

Step 1 gives the list of items that have been cached by the client, and step 2 checks whether the particular cached item has been updated when the client was in disconnection mode. Step 2 is repeated for all the cached items by the

client MH_i . The number of bits in DIV_i is C and is equal to the number of items cached by MH_i .

Once the DIV_i has been constructed, the server sends DIV_i to MH_i over the downlink channel. After downloading DIV_i , MH_i finds whether a particular cached item is valid or not. If $DIV_i[k] = 1$, then the k^{th} cached item is invalid, otherwise it is still valid. After checking for each cached item, the client will send an uplink request for all the invalid items, and the server responds by broadcasting the requested items during and following UR/RR.

As compared to UIR strategy (Cao, 2001, 2002a, 2002b, 2003), which handles disconnection by sending the ids for updated items, our strategy uses only one bit for an item, thus reducing the reconnection overheads tremendously. For our strategy, the reconnection overhead is C bits, which is very low as compared to UIR. Because of the smaller size of overheads, our strategy is also very much effective in terms of bandwidth utilization, client tuning time, and energy consumption.

An Example

Consider a database having 10 items with the last update timestamp t_i as follows:

d_i	1	2	3	4	5	6	7	8	9	10
t_i	20	16	17	13	5	6	2	9	23	19

Consider a host MH_x of cache size $C = 4$ that has cached the items with id d_1 , d_2 , d_4 , and d_7 . Let MH_x be disconnected at time 17 such that it has received the last UR at $T_1 = 15$ and wakes up at time 30.

Then:

$CSI_x =$	d_1	d_2	d_4	d_7
	↓	↓	↓	↓
$DIV_x =$	1	1	0	0

While MH_x receives DIV_x , it is interpreted as: 1st cached item where d_1 is invalid, and 2nd cached item where d_2 is invalid, whereas d_4 and d_7 are still valid. The reconnection overhead for our strategy is 4 bits.

For UIR, the overhead = number of cached items invalidated during disconnection * item id size (S_{id}). Generally $S_{id} = 32$ bits, therefore the overhead value = 64 bits.

CONCLUSION AND FUTURE DIRECTION

This chapter investigates the cache invalidation issue in a realistic mobile environment, where there exist resource-poor mobile clients, data updates, asymmetric low-quality wireless channel, and client disconnections. UR-based cache invalidation strategy has been proposed that reduces the query latency as compared to existing IR and UIR strategies. The UR strategy employs selective tuning to conserve the client's battery power. Frequent client disconnection is one of the main features in a mobile computing environment. To cater for such an environment, a disconnection information vector (DIV) based algorithm has been proposed that maintains cache consistency at the mobile client with very low overhead as compared to existing strategies. The ad hoc mode of operation, which is now available with new-generation wireless interfaces, makes possible peer-to-peer (P2P) caching in which mobile clients can access data items from the cache in their neighboring peers. Extension of the proposed strategy to peer-enabled caching is a consideration during our future research.

REFERENCES

- Barbara, D., & Imielinski, T. (1994, May 24-27). Sleepers and workaholics: Caching strategies in mobile environments. *Proceedings of the ACM SIGMOD Conference on Management of Data*, Minneapolis, MN (pp. 1-12).
- Cao, G. (2001, August 6-11). A scalable low-latency cache invalidation strategy for mobile environments. *Proceedings of the ACM International Conference on Computing and Networking (Mobicom)*, Massachusetts (pp. 200-209).
- Cao, G. (2002a). On improving the performance of cache invalidation in mobile environments. *Mobile Networks and Applications*, 7(4), 291-303.
- Cao, G. (2002b). Proactive power-aware cache management for mobile computing systems. *IEEE Transactions on Computers*, 51(6), 608-621.
- Cao, G. (2003). A scalable low-latency cache invalidation strategy for mobile environments. *IEEE Transactions on Knowledge and Data Engineering*, 15(5), 1251-1265.
- Chand, N., Joshi, R. C., & Misra, M. (2004, December). Broadcast based cache invalidation and prefetching in mobile environment. *Proceedings of the International Conference on High Performance Computing (HiPC)* (pp. 410-419). Berlin: Springer-Verlag (LNCS 3296).
- Chand, N., Joshi, R. C., & Misra, M. (2005, January 23-25). Energy efficient cache invalidation in wireless mobile environment. *Proceedings of the IEEE International Conference on Personal Wireless Communications (ICPWC)*, New Delhi, India (pp. 244-248).
- Chuang, P. J., & Hsu, C. Y. (2004, March 29-31). An efficient cache invalidation strategy in mobile environments. *Proceedings of the IEEE International Conference on Advanced Information Networking and Application (AINA)*, Fukuoka, Japan (pp. 260-263).

Jing, J., Elmagarmid, A., Helal, A., & Alonso, R. (1997). Bit-sequences: An adaptive cache invalidation method in mobile client/server environments. *Mobile Networks and Applications*, 2(2), 115-127.

Kahol, A., Khurana, S., Gupta, S. K. S., & Srimani, P. K. (2001). A strategy to manage

cache consistency in a disconnected distributed environment. *IEEE Transactions on Parallel and Distributed Systems*, 12(7), 686-700.

Tan, K. L., Cai, J., & Ooi, B. C. (2001). An evaluation of cache invalidation strategies in wireless environments. *IEEE Transactions on Parallel and Distributed Systems*, 12(8).

Chapter XI

Review of Wireless Technologies and Generations

Raghunadh K. Bhattar

Indian Space Research Organization & Indian Institute of Science, India

K. R. Ramakrishnan

Indian Institute of Science, India

K. S. Dasgupta

Indian Space Research Organization, India

V. S. Palsule

Indian Space Research Organization, India

ABSTRACT

The concept of wireless communication i.e. exchange of any type of electronic content (audio, video, data, etc.) without the use of any physical medium like cables, wires etc, is not new. It started almost a century back with “radio and telegraphs” and has made rapid progress over the period of time. Looking at the present trend, it is obvious that the field of wireless communication will continue to move far ahead than one could imagine. To encourage adoption and advancement in technology, a standardization process was required. Different standards have evolved, over the period of time, to accommodate the new technologies as they emerged. The era during which technologies are popular and are used by businesses is generally termed as Generations. This chapter presents a brief history of wireless communication and different phases of technologies and standards involved in it. A discussion of the communication technology generations not only provides an understanding of the past history of these technologies, but also creates basis for understanding their future. This chapter provides brief introduction and description of all generations, starting from the first generation mobile communications to future generation mobile communications.

INTRODUCTION

In this age of information and communications technology (ICT), there is an ever-increasing demand for almost constant delivery of information, anytime, anywhere. Knowledge is enormous and ever increasing exponentially. Both the public and business class people have become “tech-savvy.” They rely more and more on technology to overcome the barriers of distance and time, and wish to access the information instantly at any time and from anywhere. A business man wants access to urgent information while he is “on the road”, a boss wants to be in contact with his people working at remote sites, a salesperson wants continuous updates on the information related to his product, and nowadays even children demand mobile gadgets to communicate with their parents or friends. How to fulfill the dreams of such a wide range of people? The technology strives to provide solutions through development and innovations in the field of wireless communication.

Communication is a very important aspect of human life. All living beings communicate with each other in one way or another. Humans, being an intelligent, social, and restless creature, strive to find better and better ways of communicating with each other. In olden days, people were sent to far-off places as messengers. Later on, pigeons were trained as messengers for fast communication. Then fast-moving vehicles were invented and used for communication. But the dream of very fast communication became a reality only after the invention of electricity. However, humans always had a fantasy of communicating instantly, anywhere, anytime, as mentioned in many holy books. This fantasy was transformed into reality with the discovery of electromagnetic waves (Shea, 2000), which revolutionized the communication world in a holy way. Researchers are

constantly pursuing various ways to make this technical breakthrough achieve mobile communications through wireless means. As a result, in the last few decades, various standards have emerged for mobile communications (known as *generations*)—starting from the first generation (1G) with analog technology, passing through digital revolution with 2G, 3G, 4G, and so on. The standards also exist for technology in between two generations such as 2.5G technology (The Mobile Phone Directory, 2005c). As wireless technology advances, giving rise to new opportunities, customer needs increase and hence the new standards are being proposed. Now it is right time for the business community to grab this opportunity and make mobile communication an affordable luxury for everyone in the near future.

The content of the chapter is organized into different sections as follows. The first section gives a brief account of the history of wireless communications. The next section describes the development of wireless communications through wireless generations; this section covers the different technologies and main features of the 1G, 2G, 2.5G, and 3G standards, and gives a brief description of the 4G technology, including the promise it holds, requirements, and challenges. The last section concludes the chapter.

HISTORY OF WIRELESS COMMUNICATIONS

In 1867, the foundation of wireless communication was laid, when Maxwell predicted the existence of electromagnetic (EM) waves. Later in 1887, Hertz proved the same. However in 1898, a significant breakthrough came, when Marconi demonstrated the wireless telegraph by establishing wireless communication between England and France. In 1902, suc-

successful bi-directional wireless transmission was achieved. In 1914, the first voiceover radio transmission took place, and by the 1930s mobile communications were used in police cars. In the 1940s, after Armstrong demonstrated frequency modulation (FM), the majority of the police systems were converted to FM. In 1946, mobile users were connected to the public switched telephone network (PSTN), and in the 1960s, the number of mobile users crossed the one million mark. During this time, an Improved Mobile Telephone Service (IMTS) was introduced, which supported full-duplex, auto-dial, and auto-trunking facilities. In 1979, NTT/Japan deployed the first cellular communication system. Later in 1983, Advanced Mobile Phone System (AMPS) was deployed in the United States in 900MHz band and supported 666 duplex channels. During 1983-1989, Groupe Spéciale Mobile defined the European digital cellular standard called GSM. In 1991, a digital cellular phone system was introduced in the U.S., and in 1993, the code-division multiple-access (CDMA) spread-spectrum digital cellular system (IS-95) was deployed. In 1994, the GSM system was deployed in the U.S. and relabeled as the *Global System for Mobile Communications* (Shea, 2000).

DEVELOPMENT OF WIRELESS TECHNOLOGIES

First-Generation (1G) Wireless Technologies

As described in the introduction, wireless technologies have evolved in several major phases, denoted by “generations,” or “G” for short. 1G denotes first-generation mobile technology based on analog systems, deployed during the 1980s. The 1G technologies were intended to transmit only voice phone calls from wireless

handsets and hence are referred to as analog technology (Day, 2001). In 1979, Nordic Mobile Telephone (NMT) was introduced in Sweden and Norway as a commercial analog mobile system. Another popular example of first-generation cellular is Advanced Mobile Phone System (AMPS), which supported only analog voice and became operational in 1983. The voice calls were modulated and transmitted using frequency modulation (FM), and use Frequency Division Multiple Access (FDMA) for multiple access. Another system based on AMPS, known as Total Access Communications System (TACS), was intended to be used in United Kingdom, but was primarily deployed in Asia-Pacific regions with extended specifications known as ETACS (Ames & Gabor, 2000).

1G systems typically use 25MHz bandwidth each for forward and return links between the base station and handsets. These bands are split into a number of communication channels and allocated to each user (The Mobile Phone Directory, 2005a).

1G systems had many limitations. First, the calls suffered from security threats as they were sent in the unencrypted form, and hence were easy to intercept. Second, both voice quality and connection reliability was poor. Third, the handset was bulky. Last, as analog systems, 1G networks did not support wireless data. So, 1G technology became obsolete with the advent of second-generation, or 2G, cellular technologies (Day, 2001).

Figure 1. 1G standards

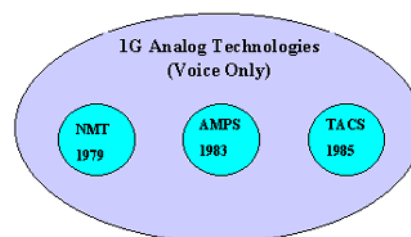
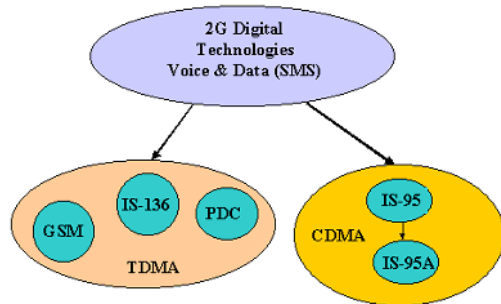


Figure 2. 2G standards



Second-Generation (2G) Wireless Technologies

To improve voice quality, network security, and call reliability, 2G networks largely depend on digital technology such as voice coding, forward error correction (FEC), and higher-order digital modulation techniques. For efficient use of the frequency spectrum, multiple access schemes such as a Time Division Multiple Access (TDMA) and Code Division Multiple Access (CDMA) are used in 2G networks. 2G standards include three TDMA standards and one CDMA standard. Due to several reasons pertaining to technical and non-technical aspects, different 2G technologies have come into existence, including the following.

Global System Mobile (GSM)

One of the most successful 2G digital technologies developed during the 1980s for cellular use, GSM technology uses both FDMA and TDMA for network access. The first GSM system used a 25MHz frequency spectrum in the 900MHz band. The 25MHz of bandwidth is divided into 124 carrier frequencies of 200kHz each using FDMA technology. Each carrier frequency is then further divided using a TDMA scheme into eight timeslots, with 577 microsec-

onds for each timeslot. Each timeslot constitutes one communication channel. Today, GSM systems operate in the 900MHz and 1.8 GHz bands throughout the world except in America, where they use the 1.9 GHz band (The Mobile Phone Directory, 2005b; Rappaport, 2002).

Interim Standard 136 (IS-136): Time Division Multiple Access (TDMA)

IS-136 supports three timeslotted users for each 30KHz radio channel and uses TDMA for multiple access. It is more efficient than 1G analog technologies. Generally, both analog and IS-136 TDMA channels can be operated on the same network. Due to this dual-mode operation, users can take advantage of both widely established analog networks and more advanced technology of IS-136 TDMA where it exists. The capacity of the IS-136 networks is approximately three times more than the analog networks, but problems due to co-channel interference may be experienced, as the data rates on the TDMA networks are higher compared to the bandwidth available (Gupta, 2005b).

Pacific Digital Cellular(PDC)

This is similar to IS-136, except in channel spacing and voice compression. It is widely used in Japan.

Table 1. 2G technologies

Mobile Standard	Average Data Throughput	Air-Interface
GSM	9.6kbps	TDMA
IS-136	9.6kbps	TDMA
PDC	9.6kbps	TDMA
IS-95	9.6kbps	CDMA
IS-95A	14.4kbps	CDMA

Interim Standard-95 Code Division Multiple Access (IS-95)

This standard is also known as *cdmaOne*, and all the users in the network occupy the entire 1.25MHz channel simultaneously using a spread spectrum technique called Code Division Multiple Access (CDMA). In contrast to FDMA or TDMA, where each user is allocated a particular frequency band or timeslot respectively, in CDMA the entire bandwidth is allocated to all users at all times. The users detect their signal with the help of orthogonal sequence codes, unique to each user in the given cell. The length of the orthogonal sequence code greatly influences the performance of the CDMA system. The bits are called “chips” and the number of bits per second in the orthogonal sequence code is generally termed as “chipping rate” in CDMA terminology. IS-95 standard supports about 60 simultaneous users in the 1.25MHz channel. The use of CDMA technology for multiple access has several advantages like improved network capacity, immunity from interference by other signals, reliable connection, improved voice quality, and network security. Further advantages of using a CDMA system are better frequency reuse, soft handoff, and so forth. Soft handoff refers to the scheme in which a subscriber is attached to two adjacent cells simultaneously during transitions from one cell to another. This way a smooth and seamless transition takes place between the cell crossing and is transparent to the user. This standard is widely deployed in North America and hence is

sometimes called North American CDMA (NA-CDMA). The IS-95 standard supports a data rate of 9.6kbps while the slightly improved version, IS-95A, supports data rates of 14.4kbps (Gupta, 2005a; Rappaport, 2002).

2G is the first set of wireless standards to employ digital modulation and advanced digital signal processing in both handset and base station for air interface. Compared to first-generation analog technologies, all 2G technologies offer minimum three-fold increase in bandwidth efficiency to meet the requirements of increasing demands of the cellular users. The data rates around 9.6kbps are supported for each user in most of the 2G standards, but these data rates are generally not sufficient for fast Internet applications. This is because 2G standards use circuit-switched technology for data transmission, and one full voice channel is dedicated for each connection. However, with the circuit-switched approach, one popular data service offered by GSM is Short Messaging Service (SMS), which is very popular with GSM (Rappaport, 2002).

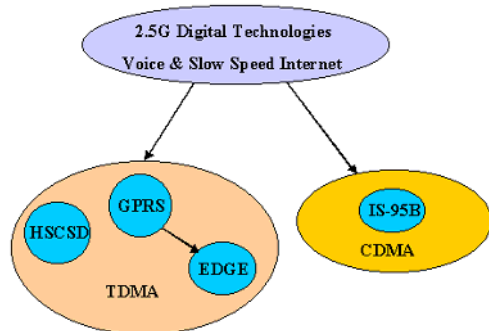
2.5G Technology

To support fast and data-intensive Internet applications, new standards have been proposed with higher data rates. The aim of these proposed standards is to have backward compatibility with 2G standards and to build upon the existing 2G technologies, as the 2G standards are widely deployed. These new stan-

Table 2. The major upgrade paths for 2.5G technology

Technology Upgrade to 2.5G	Base 2G Technology
High-Speed Circuit-Switched Data (HSCSD)	GSM
General Packet Radio Services (GPRS)	GSM & IS-136
Enhanced Data Rates for GSM Evolution (EDGE)	GSM & IS-136
IS-95B	IS-95A

Figure 3. 2.5G standards



dards are termed as 2.5G technology, and allow existing 2G handsets and base stations to be modified with software upgrades and add-on cards to support higher data rates for Internet applications such as e-mail, Web browsing, mobile commerce, and location-based mobile services. The 2.5G standards also support Wireless Applications Protocol (WAP). The WAP is a new Web browsing language which supports compressed Web pages to be viewed on small portable handheld devices. Before the introduction of WAP, NTT DoCoMo in Japan deployed a proprietary wireless technology, called iMode, on its PDC network in 1988. iMode has the capability to support interactive Internet browsing with color graphics, online games, and so forth, with only 9.6kbps bit rates on its PDC network. Similarly, a wide range of 2.5G standards have been deployed to allow each of the popular 2G technologies such as GSM, IS-136, and cdmaOne (IS-95) to be upgraded for faster Internet access (Rappaport, 2002; Continuous Computing, 2005) as mentioned in Table 2.

High Speed Circuit-Switched Data (HSCSD)

This allows individual users to use consecutive timeslots in order to obtain high-speed data

access on the GSM network, unlike limiting each user to only one specific timeslot as in the 2G GSM standard. HSCSD provides application data rates up to 14.4kbps by using flexible error control coding, as compared to the original GSM specification of 9.6kbps. Further, it is possible to obtain a raw bit rate up to 57.6kbps for each individual user by using four consecutive timeslots (Rappaport, 2002).

General Packet Radio Services (GPRS)

This is a packet-based data network on dedicated GSM or IS-136 radio channels, and suitable for non-real-time Internet applications and asymmetric Web browsing. Due to this, GPRS networks can be shared by multiple users in contrast to HSCSD networks, where the channel is dedicated to a single user because of the circuit-switched mode. As a consequence, the GPRS can support a larger number of users than HSCSD, but in bursty mode. The modulation schemes for GPRS are the same as that of the 2G TDMA standard, but they use a totally different air interface to improve the packet data handling and access. In GPRS, applications are required to include their own error correction schemes to their payload data. GPRS can achieve data rates up to a maximum

Figure 4. Upgrade from 2G to 2.5G

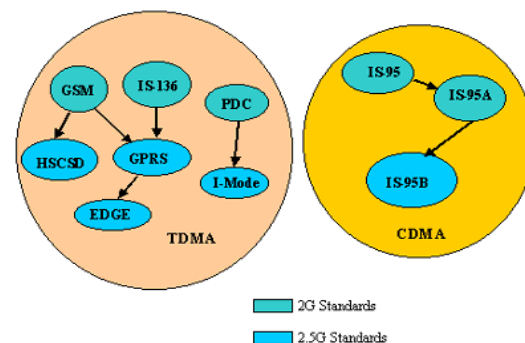


Table 3. 2.5 technologies

Mobile Standard	Average Data Throughput	Peak Data Rate	Air-Interface
HSCSD	14.4kbps	57.6kbps	TDMA
GPRS	50kbps	171.2kbps	TDMA
EDGE	384kbps	> 1Mbps	TDMA
IS-95B	64kbps	115.2kbps	CDMA

171.2kbps by using all available eight timeslots of the GSM radio channel, but drops to 40-60kbps when more users access the network or network conditions becomes poor. The GPRS technology provides “always-on” capability to user handsets, which are automatically tuned to a dedicated GPRS radio channel and particular timeslots when required (General Packet Radio Service, 2004; Rappaport, 2002).

Enhanced Data Rates for GSM Evolution (EDGE)

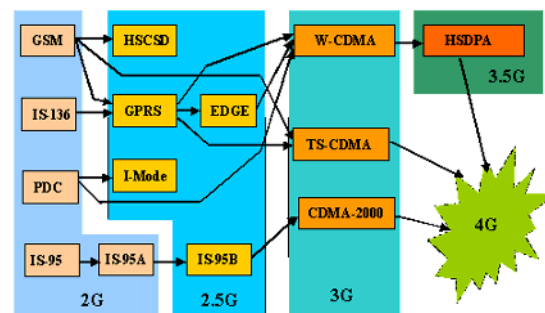
In addition to GSM’s standard Gaussian Minimum Shift Keying (GMSK) modulation, EDGE uses 8-PSK (8-Phase Shift Keying) modulation for higher data rate support. EDGE supports nine different air interface formats, known as Multiple Modulation and Coding Schemes (MCS), with different levels of error control. MCS can be personalized for each individual user. In fact, depending on the network conditions, the best MCS settings are adaptively determined for each individual user. This adaptive capability to select the “best” air interface is called incremental redundancy. In incremental redundancy, acceptable level of quality is quickly attained and maintained with minimum resource allocation, with the help of the feedback from the subscriber to the base station. Typically EDGE provides raw data rates around 384kbps for a single user on a GSM channel. It is feasible for EDGE technology to provide data rates up to several megabits per second using higher-order modulation, multi-carrier transmission techniques, and so forth (Rappaport, 2002). EDGE is also sometimes referred to as

2.75G technology, due to its high data rate support compared to other 2.5G technologies. These 2.5G upgrade options provide significant speed improvements in Internet access and also make handset Internet usage possible (Continuous Computing, 2005).

IS-95B for 2.5G CDMA

The IS-95B is an upgrade path for IS-95 and represents 2.5G technology. IS-95B provides high data rates to individual users, by combining multiple orthogonal (coded) user channels together on a common CDMA radio channel. IS-95B supports both packet and circuit-switched data access. IS-95B allows up to eight CDMA channels to be combined to provide data rates up to 115.2kbps per user. But the slotting techniques used in the air interface limits the usable data rates to around 64kbps for a single user. IS-95B adopts some hard handoff procedures to maintain better link quality (Rappaport, 2002).

Figure 5. Upgrade path from 2G to 3G and beyond



Even though the 2.5G standards offer Internet services, the data throughput rates are not sufficient for high-speed broadband Internet and multimedia access. Also 2.5G standards do not offer many services, and also services are not flexible enough to users. Hence, the need arose for next-generation technologies, termed as 3G technology, to meet these demands.

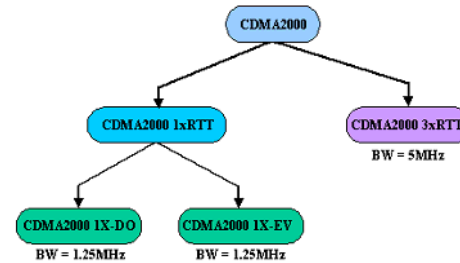
Third-Generation (3G) Wireless Technologies

3G wireless technologies promise a wide range of wireless data access and services in a manner “never seen before.” 3G systems offer wide bandwidth, high transmission speeds, huge network capacity, and global roaming. This enables high-speed Internet access in the order of several megabits per second, Voice over IP (VoIP), fax, e-mail, audio and video streaming, multimedia services on demand, and videoconferencing, to name a few. All these are possible with a single mobile handset in an “always-on” access fashion.

Initially, the International Telecommunications Union (ITU) planned to establish a common wireless communication standard, which will be applicable to all countries in the world using a global frequency band in the range of 2000MHz frequency. This plan is known as International Mobile Telephone 2000 (IMT-2000), and up to a certain extent was successful in bringing the people together for active discussion and debate. But unfortunately, it failed in its aim of unified standard, as the wireless community was divided into two groups. One group belongs to TDMA-based GSM, IS-136, and PDC, and the other group belongs to CDMA. These two groups are known as 3GPP (3G Partnership Project for Wideband CDMA standard) and 3GPP2 (3G Partnership Project for CDMA2000 standard).

The ITU specifies minimum data rates that should be supported by 3G standards, depend-

Figure 6. CDMA variations



ing on the speed of the mobile user. Data rates of 144kbps should be supported for the user moving with a speed of more than 120kmph in outdoor environments. Similarly, 384kbps should be supported for pedestrian users moving with speed less than 120kmph and 2Mbps data rates for stationary or campus moving environments (Hughes, Kelly, Levine, & Pillsbury, 2001).

Most of the 3G standards maintain the backward compatibility with 2G and 2.5G standards in their group. The backward compatibility implies that the 2G and 2.5G users can avail the services offered by 3G networks in full or with limited options, with no change or with some minor modifications to their handsets. However the base station equipment of 2G or 2.5G may need a complete change to facilitate this backward compatibility.

There are several technologies contended for 3G solutions. Three of these technologies (also coming under the banner of IMT-2000) are described here. They are:

- W-CDMA (UMTS),
- CDMA2000, and
- TS-SCDMA.

W-CDMA (UMTS)

Wideband-CDMA or W-CDMA, also known as Universal Mobile Telephone System (UMTS), uses the CDMA technology with 5MHz wide bandwidth for air interface, hence

the name Wideband-CDMA. The network architecture of W-CDMA is based on GSM and EDGE, and is backward compatible with TDMA-based 2G and 2.5G technologies. It implies that both the GSM and W-CDMA technologies can share the same network. Further the backward compatibility feature allows handover between these technologies possible with dual-mode handsets. But the hardware and signaling mechanisms are different, as GSM uses TDMA, whereas W-CDMA uses CDMA technology for radio access. Also, bandwidth allocation for each user is different in both the systems (Ericsson, 2001).

The new CDMA air interface provides the much-needed additional capacity and bandwidth. The UMTS air-interface standards provide wireless services to a wide range of devices and equipment such as computers, mobile handsets, multimedia devices, and so forth. All these devices can use the same network at the same time for Internet and multimedia data services at anytime, anywhere, with “always-on” access capability of 3G networks. W-CDMA wireless services are packet based and support data rates up to 2Mbps per user. These data rates allow data-intensive multimedia, streaming of live audio and video, and high-speed Internet services (Rappaport, 2002).

CDMA2000

CDMA2000 is an “evolutionary” 3G standard, which means that the technology evolves in a gradual manner, keeping the new standards backward compatible with existing cdmaOne standards. This results in seamless migration from 2G and 2.5G technologies to 3G technologies. Further, the same network will support different standards and different handsets belonging from 2G to 3G standards. Thus the CDMA2000 standards are backward compatible with IS-95, IS-95A, and IS-95B standards.

Since CDMA2000 is an evolutionary standard, the upgrade path is smoother and cost effective for both the service providers and customers.

The CDMA2000 standard is being developed by the U.S.-based Telecommunications Industry Association (TIA) in association with the 3GPP2 working group. The standard is evolving through different phases to satisfy the large variety of customer needs and wide range of applications. Some of these variations include:

- CDMA2000 1xRTT,
- CDMA2000 3xRTT,
- CDMA2000 1xEV-DO, and
- CDMA2000 1xEV-DV.

The RTT stands for Radio Transmission Technology and is specified by the IMT-2000 standardization body. 1x indicates that the bandwidth is one times that of the original cdmaOne channel and is modulated on a single carrier. Similarly, 3x indicates that the bandwidth is three times that of the original cdmaOne channel and uses multi-carrier modulation with three carriers.

The first 3G CDMA air interface, also called CDMA2000-1xRTT or simply CDMA2000 1X, operates on the same bandwidth of 1.25MHz per channel, which was originally specified for 2G CDMA standard. This ensures that the backward compatibility with cdmaOne standards is maintained. For voice communications, CDMA2000 1X supports packet data rates of 307kbps peak and typical packet data rates about 144kbps per user. The data rates vary depending on several network conditions such as mobility of the user, the propagation conditions, and network load. These data rates are about twice that of cdmaOne. The higher data rates in CDMA2000 1X are obtained due to improved base band signaling and chipping rates. The existing RF equipment is sufficient to obtain these higher data rates, as the

Table 4. 3G and 3.5G technologies

Mobile Standard		Peak Data Rate	Air-Interface
W-CDMA		2Mbps	CDMA
CDMA2000	1xRTT	307kbps	CDMA
	3xRTT	> 2Mbps	CDMA
	1xEV-DO	2.4Mbps	CDMA
	1xEV-DV	~ 5Mbps	CDMA
TS-CDMA		2Mbps	TDMA+CDMA
HSDPA		10Mbps	CDMA

CDMA2000 1X operates on the same bandwidth as that of cdmaOne. Only changes required may be in software and base band hardware (Rappaport, 2002).

To meet the higher data rate requirements of matured 3G standards, bandwidth needs to be increased. The increase in the bandwidth can be achieved through clubbing the adjacent radio channels of cdmaOne and using multi-carrier techniques. CDMA2000 3xRTT uses these techniques by combining three adjacent 1.25MHz bandwidth of cdmaOne channels to yield a bandwidth of around 5MHz (including guard band). This bandwidth improvement gives the packet data rates in excess of 2Mbps per user depending upon network conditions and vehicle speed. CDMA2000 3xRTT can be compared with W-CDMA in many aspects such as bandwidth, data rates, and so forth, and differs in many areas like chip rates, frame lengths, and base station synchronization (Rappaport, 2002).

CDMA2000 1xEV is an evolutionary upgrade for CDMA2000 and is part of IMT-2000. The CDMA2000 1xEV standard provides two options for accessing only data (CDMA2000 1xEV-DO) or both data and voice (CDMA2000 1xEV-DV).

The CDMA2000 1xEV-DO option provides data rates of about 2.4Mbps and supports data traffic only. No voice communication is supported. This relaxes the minimum latency requirements and end-to-end delay restrictions on the radio channel to improve the data rates.

This option is useful and intended for high-speed wireless Internet access. This mode of operation is not backward compatible with CDMA2000. Even though the peak data rates are in excess of 2.4Mbps, the real data rates may vary depending on the vehicle speed, network load, and propagation conditions, and may drop to a few kilobits per second.

CDMA2000 1xEV-DV will support peak bit rates far above 2Mbps, while the usable data rates available to the user in practical conditions may be limited to less than 2Mbps. CDMA 1xEV-DV will be backward compatible with CDMA2000 and promises to offer wireless data, voice, high-speed Internet access, videoconferencing, multimedia access, streaming of audio and video, online games, and so forth (Texas Instruments, 2005; Rappaport, 2002).

TD-SCDMA

Time-Division Synchronous Code Division Multiple Access (TD-SCDMA) was proposed by the Chinese Academy of Telecommunications Technology (CATT) and Siemens Corporation, and adopted by ITU as one of the 3G options. TD-SCDMA uses the existing GSM infrastructure with additional upgrade tools to obtain higher data rates to make it a candidate technology for 3G standards. TD-SCDMA uses both TDMA/TDD and CDMA together in synchronous mode to support both circuit-switched

data and packet-switched data. TDMA divides each frequency into number of timeslots, which can be flexibly assigned to different users, and the TDD (Time Division Duplex) mode allows both uplink and downlink to operate in the same frequency band but at different timeslots. Further, CDMA technology enables more users to access the network at the same time in each timeslot.

The TD-SCDMA radio channel operates on 1.6MHz bandwidth and supports data rates up to 2Mbps. TD-SCDMA uses several new technologies such as joint detection, dynamic channel allocation, smart antennas, and mutual terminal synchronization. These techniques improve the transmission capacity, inter-cell interference, and system cost, and eliminate the expensive soft handover requirements (Rappaport, 2002).

High Speed Downlink Packet Access (HSDPA)

HSDPA is an upgraded path to W-CDMA and offers data rates up to 8-10Mbps in a 5MHz bandwidth radio channel. HSDPA archives these data rates using Adaptive Modulation and Coding (AMC) techniques, Multiple-Input Multiple-Output (MIMO), Hybrid Automatic Request, improved cell search, and receiver design. This is also sometimes referred to as 3.5G technology, because HSDPA is an intermediate packet-based data service solution before leaping into 4G technologies (UMTS, 2003).

Fourth-Generation (4G) Wireless Technologies—Future Mobile Communication Technology

Initially, voice communication was only possible with 1G analog technologies. The world was happy when digital technology emerged in a big way and became realizable, miniaturized, and cheaper; the world turned to digital tech-

nology for wireless communications. In 2G standards, in addition to voice communication, SMS services were added. Then arose the demand for wireless Internet services. 2.5G technology provided that. But then arose the demand for fast Internet services, live video and audio streaming, online games, interactive multimedia, and much more. 3G was evolved to address these problems, and it does offer most of these services, but in a limited way. As the data rate support of 3G is around 2Mbps, all these desired services cannot be available at the same time at the desired speed.

The next-generation technology, the 4G, has ambitious goals as well as challenging research issues and above all never ending user expectations. The research activities to achieve these goals have already started, and the standardization process and commercial availability are expected by the year 2010. Typical data rate throughputs of about 20Mbps are targeted, with the users moving at an average speed of 200kmph. The peak data rates can go up as high as 100Mbps. 4G is expected to operate in 2-8GHz bandwidth. To achieve these targets, a large amount of research efforts are needed in a wide range of fields (LeFevre & Okrah, 2001).

4G standard will integrate almost all different networks, including heterogeneous networks available today and those that will appear in the future. Users can communicate and roam freely between these networks and stay connected through single mobile handset. There are exciting possibilities waiting with the integration of networks such as Personal Area Networks (PANs), Bluetooth, and so forth, with 4G networks.

The network architecture needs to be improved for accommodating inter-portability and seamless handover, not only among the same network but also between the two different networks when the user moves. For example, 4G standards should support handover between

the CDMA network and GSM network or vice versa with the same handset. This is known as vertical handoff. To support the same handheld device to operate on different networks with different air-interfaces, technologies like Software Defined Radio are required. To support a huge number of users, the network capacity needs to be improved tremendously. Also network architecture should support a wide range of bit rates from 10Mbps to 100Mbps and wide bandwidth ranging from 2GHz to 8GHz frequency.

Multi-carrier CDMA (MC-CDMA) is a prime candidate for 4G air-interface. The effects of MAI (Multiple Access Interference) and ISI (Inter Symbol Interference), which are critical for better BER (bit error rate) performance, can be reduced considerably with MC-CDMA. MC-CDMA is a combination of CDMA and OFDM (Orthogonal Frequency Division Multiplexing) (Krishnakumar, n.d.).

Adaptive Modulation and Coding (AMC) schemes are also considered for 4G technologies in view of high data rate support required at high vehicle speed. AMC changes its modulation and coding schemes dynamically according to the varying channel conditions and adopts accordingly (Lu, 2003). Smart antennas will be used to improve the signal strength at the mobile handset. This requires sophisticated base band signal processing and requires DSPs, FPGAs, and ASICs for implementation (LeFevre & Okrah, 2001).

Some of the research issues involved in development of 4G technologies are listed below (Krishnakumar, n.d.).

- Low Cost
- Better Quality of Service

Though both goals appear to be mutually contradicting, the research should aim at producing 4G systems at low cost with better quality of service, to make the 4G technology

affordable to everybody and preferred by everybody.

- Bandwidth Efficiency
- High Bandwidth Support
- High-Frequency Range Operation

Considerable efforts should be made so the maximum use of available bandwidth is used in the most efficient way to accommodate more and more users, and to be able to provide as many services as promised by 4G. There is a need to use larger bandwidth and a higher side of the spectrum to accommodate new services, more users, high-speed communications, and multimedia services. A variety of problems will be encountered in high-frequency range. For example, in high-frequency bands, the signal becomes distorted or attenuated due to rain and other atmospheric conditions. Powerful signal processing techniques are required to overcome these problems.

- Increased Data Rates

High data rate operation is a real challenge for communication people, as the data need to be delivered at varying channel conditions and at high vehicle speeds. The sophisticated signal processing techniques, use of smart antennas, and high-performance receiver technologies will significantly enhance the data rates with better quality of service.

- High-Speed Internet Services
- IP-Based Data Access
- Increased Data Throughput
- Increased Network Capacity

As proposed, 4G wireless communications will be IP based, and should support high-speed Internet access. So the present Internet protocols should be extended or modified to a wireless environment. It is expected that the num-

ber of 4G users will be extremely large, hence the network capacity and data throughput should be increased many times over. Great research effort is needed in this area.

- Improved Multimedia Services
- Smooth Streaming of Video and Audio
- Browser Technologies

In present times, multimedia (MM) communication has become very popular. People prefer visual communication to voice communication and Multimedia Messaging Service (MMS) to Short Messaging Service (SMS). Camera-equipped mobile phones are now common. Hence the wireless technologies have to support MM services. The multimedia content requires huge bandwidth for transmission, and a high amount of memory for storage and display. Sophisticated signal processing techniques and extremely fast hardware and software are required to meet these challenges. Researchers should also look for improved browser technologies to effectively handle MM content and provide simple and conventional maneuverability. Also the browsers should be capable of handling smooth audio and video streaming for live visual communications.

- Universal Access
- Global Roaming
- Portability Between Different Devices
- Handoff Techniques
- Software Defined Radio
- Tight Network Security

Universal access and global roaming refer to the capability of a mobile user to access the wireless services from anywhere with his single handset irrespective of the type of the services provided by the local service provider. Hence, research efforts are required to make the user handset support different mobile technologies

such as GSM, CDMA, and so forth, as well as configurable automatically using techniques such as software defined radio. Much research needs to focus on better vertical handoff techniques in addition to improved horizontal handoff techniques. In the context of universal access, improved and hack-free network security becomes an issue.

- Efficient Multiple Access Techniques
- Adaptive Modulation and Coding
- Higher-Order Modulation Schemes
- Development of New Algorithms

These are the core issues to be tackled by communication researchers. A primary area of 4G research includes bandwidth-efficient multiple access techniques, adaptive techniques in coding and modulation, and higher-order modulation schemes. To achieve many ambitious goals of 4G, innovative techniques need to be investigated in the above areas and related fields. These efforts may result in new algorithms to improvement the quality of service, increased data throughput and spectral efficiency, and so forth.

- Source Coding
- Channel Coding
- Joint Source Channel Coding
- Error Resilient Coding

Source coding refers to removing the redundancy in sensor-generated data such as audio, speech, video, and data; this will help in reducing the data to be transmitted over wireless channels. Redundancy removed data is more susceptible to channel noise, and some controlled redundant data need to be added to detect and correct the possible errors introduced by the channel. MPEG-4 (Richardson, 2003) is a state-of-the-art source coding technology for multimedia data, and Low Density

Parity Check (LDPC) codes (Gallager, 1962) are for channel coding. There are several research issues in these two areas, such as computational complexity, optimality, and efficient algorithms for implementation. Also, much focus is on jointly attacking the source coding and channel coding to yield maximum benefits. Error resilient coding is one example of joint source channel coding generally used in image and video coding.

- Reliability
- Fault Tolerance Systems

Above all, reliable and robust systems are required so that users can use the system all the time with minimum breakdown. This is an important factor in keeping the services preferred by everybody.

- Billing System

Last but not the least, real efforts are also required for a proper billing system. Keeping in mind that the services are available universally and under heterogeneous systems, proper tracking of the user and assessment of his consumption proportionately are a real challenge.

This is a never ending list. There are many research issues not mentioned above. The issues in antennas, power control, battery life, VLSI and ASIC technologies, and device technologies are but a few.

As we can see, 4G has many ambitious goals and exciting promises. This is the scenario even before we all experience 3G technology. The GPRS and EDGE of 2.5G are just penetrating throughout the world for Internet applications. 3G technologies are yet to be deployed in many parts of the world. 4G technologies may have a greater impact on the wireless world than 3G technologies.

Beyond 4G

Talking about 5G technologies may be premature at this stage. 5G technologies will address issues that we cannot even think about today. With 5G, we may be able to download a full movie on our handset and view it while we are on the move. It may be possible to have the cable network on our handset including HDTV. Meteorological and weather forecast departments may use 5G to download satellite images directly onto their mobile sets, and analyze the data and give the prediction instantaneously at anytime and anywhere. What is next?

A fictionist may think of transmitting your inner feelings or what you are thinking in your mind to a remote client. A food lover may want to smell through his handset the delicious food located miles away from him. A nature lover loves not only to view the beautiful scenes, but may also want to experience the fragrance of the flowers located at the remote location on his handset. All these may be perhaps addressed in 6G or XG technology. Let us wait and see!

CONCLUSION

With the advent of 2G, the era of 1G analog systems is over for all practical purposes, leaving some isolated pockets where the analog systems still exist. 2G mobile networks are widespread throughout the world and have become a part of mankind today. To cater for more services than just voice communication and SMS services, 2G technologies are being replaced with 2.5G, which offers Internet and e-mail service in addition to voice communication. In this chapter we have briefly introduced the different 1G technologies and described different 2G technologies such as GSM, IS-136, PDC, IS-95, and IS-95A. Further the

technology and compatibility issues of 2.5G upgrade paths such as GPRS, EDGE, HSCSD, and IS-95B standards were also discussed. Next, the 3G technologies, which are just hitting the market, are presented. Three important 3G technologies, W-CDMA (UMTS), CDMA2000, and TS-SCDMA, which are endorsed by IMT-2000, are explained in detail. The future-generation technology, 4G, which has a great promise and challenging research issues is discussed in some length. The user requirements and expectations, technological challenges, and research issues are then highlighted. The chapter concludes with a brief note on 5G technologies.

ACKNOWLEDGMENT

The authors gratefully acknowledge the efforts taken by Mrs. Sonal Shah for careful proof-reading of the chapter.

REFERENCES

- Ames, P., & Gabor, J. (2000). The evolution of third-generation cellular standards. *Intel Technology Journal*, 4(2). Retrieved October 7, 2005, from ftp://download.intel.com/technology/itj/q22000/pdf/art_6.pdf
- Continuous Computing. (2005). *Trillium third-generation (3G) wireless software*. Retrieved October 7, 2005, from <http://www.ccpu.com/products/papers/trillium3g.html>
- Day, B. (2001). *What is 1G wireless technology? How does it relate to J2ME?* Retrieved January 5, 2005, from <http://www.jguru.com/faq/J2ME>
- Ericsson. (2001). *White paper: Ericsson radio systems AB: Basic concepts of WCDMA radio access network*. Retrieved February 22, 2005, from http://www.ericsson.com/products/white_papers_pdf/e207_whitepaper_ny_k1.pdf
- Gallager, R. G. (1962). Low density parity check codes. *IRE Transactions on Information Theory*, IT-8(January), 21-28.
- General Packet Radio Service. (2004). Retrieved January 11, 2005, from <http://www.mobilein.com/gprs.htm>
- Gupta, P. (2005a). *Mobile wireless communications today: Code Division Multiple Access (CDMA) technology (IS-95) (cdmaOne)*. Retrieved February 22, 2005, from <http://www.wirelessdevnet.com/channels/wireless/training/mobilewirelesstoday4.html>
- Gupta, P. (2005b). *Mobile wireless communications today: Time Division Multiple Access (TDMA) IS-136 technology*. Retrieved January 11, 2005, from <http://www.wirelessdevnet.com/channels/wireless/training/mobilewirelesstoday3.html>
- Hughes, G., Kelly, J., Levine, D., & Pillsbury, D. (2001). *Wireless' third generation, will we see it and will it be worth the wait?* Retrieved October 7, 2005, from <http://www.ranjaygulati.com/new/research/WIRE-3.pdf>
- Krishnakumar, G. (n.d.). *Challenges in 4G wireless communication*. Retrieved February 25, 2005, from <http://stuweb.ee.mtu.edu/~gkrishna/final%20report.pdf>
- LeFevre, M., & Okrah, P. (2001). Fundamental changes required in modulation and signal processing for 4G. *Communications Systems Design Magazine*, (July). Retrieved March 3, 2005, from http://www.mobileinfo.com/3G/4G_CommSystemArticle.htm
- Lu, W. W. (2003). 4G mobile research in Asia. *IEEE Communications Magazine*, 41(3), 104-106.

Rappaport, T.S. (2002). *Wireless communications principles and practice* (2nd ed.). Singapore: Pearson Education.

Richardson, I. E. G. (2003). *H.264 and MPEG-4 video compression video coding for next-generation multimedia*. London: John Wiley & Sons.

Shea, J. (2000). *Brief history of wireless communications*. Retrieved January 5, 2005, from <http://wireless.ece.ufl.edu/~jshea/eel6509/misc/history.html>

Texas Instruments. (2005). *CDMA: CDMA2000 1xEV-DV*. Retrieved February 22, 2005, from <http://focus.ti.com/general/docs/wtbu/wtbuproducontent.tsp?templateId=6123&navigationId=11963&path=templatedata/cm/product/data/cdmaevdv>

The Mobile Phone Directory. (2005a). *1G—first generation networks*. Retrieved January 16, 2005, from http://www.mobile-phone-directory.org/Technology/1G_-_First_Generation/

The Mobile Phone Directory. (2005b). *2G—second generation networks*. Retrieved January 16, 2005, from http://www.mobile-phone-directory.org/Technology/2G_-_Second_Generation/

The Mobile Phone Directory. (2005c). *Introduction to mobile communications*. Retrieved January 16, 2005, from <http://www.mobile-phone-directory.org/Technology/Introduction/>

UMTS: HSDPA in W-CDMA. (2003). Retrieved February 23, 2005, from <http://www.umtsworld.com/technology/hsdpa.htm>

Chapter XII

Orthogonal Complex Quadrature Phase Shift Keying (OCQPSK) Spreading for 3G W-CDMA Systems

Shailendra Mishra

Dehradun Institute of Technology, India

Nipur Singh

Dehradun Institute of Technology, India

ABSTRACT

A variety of digital modulation techniques are currently being used in wireless communication systems. In 3G (third generation) spread-spectrum systems, such as W-CDMA (3GPP) and cdma2000 (3GPP2), the handset can transmit multiple channels at different amplitude levels. Modulation schemes such as OQPSK or GMSK do not prevent zero-crossings for multiple channels and are no longer suitable. There is a need for a modulation format or a spreading technique that can accommodate multiple channels at different power levels while producing signals with low peak-to-average power ratios. OCQPSK (Orthogonal Complex Quadrature Phase Shift Keying) has been proposed as the spreading technique for W-CDMA and cdma2000. OCQPSK is a complex spreading scheme that is very different from the modulation formats commonly used until now. The objective of this Chapter is to provide an overview of OCQPSK and explain how to start making modulation quality measurements on the reverse link (uplink) of 3G spread-spectrum systems. This chapter starts with the basic structure of the reverse link (uplink) for W-CDMA and cdma2000 with no scrambling, and explains the transition through complex scrambling to OCQPSK. The block diagrams shown are generic block diagrams for OCQPSK that are not particular to either W-CDMA or cdma2000. The chapter then describes: (1) why complex scrambling is used and how it works, and (2) why OCQPSK is used and how it works. Finally, this chapter provides how to measure modulation quality on the reverse link of 3G systems and a complete downlink physical layer model showing various results of BER and BLER calculation and also various time scopes and power spectrums.

INTRODUCTION

W-CDMA in Third-Generation Systems

Communications systems based on Spread Spectrum (SS) have been in use for decades, but most of them until the last decade were implemented in military systems because of their inherent anti-jamming and low probability-of-intercept features. In the late 1980s, the use of Direct Sequence (DS) and Frequency Hopping (FH) techniques for SS became of interest for commercial use in cellular-type communications. Code Division Multiple Access (CDMA) and Wideband Code Division Multiple Access (W-CDMA) systems appear to be the first such systems to be proposed. The wideband system was envisioned as an overlay of conventional microwave signals with a spread bandwidth (SBW) larger than the currently projected W-CDMA systems (Ojanpera & Prasad, 1998; Dahlman, Gudmundson, Nilsson, & Skold, 1998).

Analog cellular systems are commonly referred to as first-generation systems. The main first-generation standards are AMPS, TACS, and NMT. The digital systems currently in use, such as GSM, PDC, cdmaOne (IS-95)—and US-TDMA (IS-136)—are second-generation systems. These systems have enabled voice communications to go wireless in many of the leading markets, and customers are increasingly also finding value in other services such as text messaging and access to data networks, which are starting to grow rapidly.

Third-generation systems are designed for multimedia communication; with them, person-to-person communication can be enhanced with high-quality images and video, and access to information and services on public and private networks will be enhanced by the higher data rates and new flexible communication capabilities of third-generation systems. This, together

with the continuing evolution of the second-generation systems, will create new business opportunities not only for manufacturers and operators, but also for the providers of content and applications using these networks (3GPP TS 25.213 V2.3.0, 1999). With the introduction of the third generation (UMTS/IMT-2000), second-generation systems will create new business opportunities not only for manufacturers and operators, but also for the providers of content and applications using these networks (3GPP TS 25.213 V2.3.0, 1999). With the introduction of the third generation (UMTS/IMT-2000), second-generation capabilities (voice and low/medium rate data) are extended, adding multimedia capabilities to second-generation platforms such as support for high bit rates and introduction of packet data/IP access (3GPP TS 25.211 V5.0.0, 2002-2003).

In the standardization forums, W-CDMA technology has emerged as the most widely adopted third air interface. Its specification has been created in 3GPP (Third Generation Partnership Project), which is the joint standardization project of the standardization bodies from Europe, Japan, Korea, the United States, and China. Within 3GPP, W-CDMA is called UTRA (Universal Terrestrial Radio Access) FDD (Frequency Division Duplex) and TDD (Time Division Duplex), the name W-CDMA being used to cover both FDD and TDD operation (3GPP TS 25.211 V5.0.0, 2002-2003; 3GPP TS 25.201 V5.0.0, 2001-2002; 3GPP TS 25.101 V5.2.0, 2002-2003).

BASIC STRUCTURE OF THE REVERSE LINK (UPLINK) OF 3G SYSTEMS

Unlike 2G systems, in 3G systems such as W-CDMA and cdma2000, the mobile station can transmit more than one channel. The different channels are used for control purposes or to send voice and/or high-speed data. For ex-

ample, for W-CDMA the basic uplink signal comprises one Dedicated Physical Data Channel (DPDCH) and one Dedicated Physical Control Channel (DPCCH) (Dahlman et al., 1998). The DPCCH carries control information, such as an embedded Pilot that allows for synchronous detection. The DPDCH carries voice or data. Optionally, more DPDCHs may be added to support higher data rates. In the case of cdma2000, the mobile transmits a Reverse Pilot (R-Pilot) channel to allow the base station to perform synchronous detection. Additional channels, such as the Reverse Fundamental Channel (R-FCH) and Reverse Supplemental Channels (RSCHs), are used to send voice and high-speed data, respectively (3GPP TS 25.214 V5.0.0, 2002-2003). Although the names and frame structure of the channels are currently different for W-CDMA and cdma2000, the basic block diagrams of the reverse links (uplinks) are very similar. The block diagrams used in this chapter do not exactly correspond to W-CDMA or cdma2000. They are generic block diagrams that explain the evolution of the 3G reverse link structure from basic complex scrambling to OCQPSK. For both W-CDMA and cdma2000, the channels are I/Q multiplexed. In the case of transmitting only two channels, one of the channels (DPDCH or R-Pilot) is applied to the I path and the other channel (DPCCH or R-FCH) is applied to the Q path. Additional high data rate channels are combined alternatively on the I or Q paths. Each channel is spread by a different orthogonal code. The different channels can be at different power levels. If no other spreading or scrambling is applied, then I and Q signals are directly filtered and applied to the I/Q modulator. In that scheme, channels with different power levels result in different amplitudes for I and Q, which can produce strange constellations. For example, in the case of only two channels, unequal power levels result in a rectangular 4-QAM constellation. In general, equal

distribution of powers between the axes is desired, especially at the receiver, and equal powers allow for symmetry between the I and the Q paths.

Complex Scrambling

In the reverse link of cdmaOne systems, both the I and Q signals are scrambled with a PN (Pseudo-Noise) sequence prior to modulation. PN sequences from different users are uncorrelated to each other, which allow the base station to recover the signal from the appropriate user with minimum interference from others. Instead of traditional scrambling, complex scrambling has been proposed for the reverse link (uplink) for both W-CDMA and cdma2000 systems. In addition to providing differentiation among users, complex scrambling fixes the unequal distribution of powers by continuously rotating the constellation and thereby distributing the power evenly between the axes. Thus, the receiver does not have to deal with different power loads for the I and Q paths. Figure 1 shows a block diagram of the reverse link (uplink) with basic complex scrambling. After complex scrambling, the resulting I and Q base band signals are filtered and used to modulate the carrier.

Figure 2 shows the block diagram for complex scrambling. Mathematically, complex scrambling performs the multiplication of two complex signals: the complex data signal, which has already been spread into chips ($I_{\text{chip}} + jQ_{\text{chip}}$), and the complex scrambling signal ($I_s + jQ_s$). To verify this, derive the expression for the final I and Q signals by following the block diagram:

1. $I = I_{\text{chip}} - I_s - Q_{\text{chip}} - Q_s$
2. $Q = I_{\text{chip}} - Q_s + Q_{\text{chip}} - I_s$

The data signal, scrambling signal, and resulting base band signal are not really complex signals. However, the resulting I and Q signals

Figure 1. Basic reverse channel structure of 3G systems

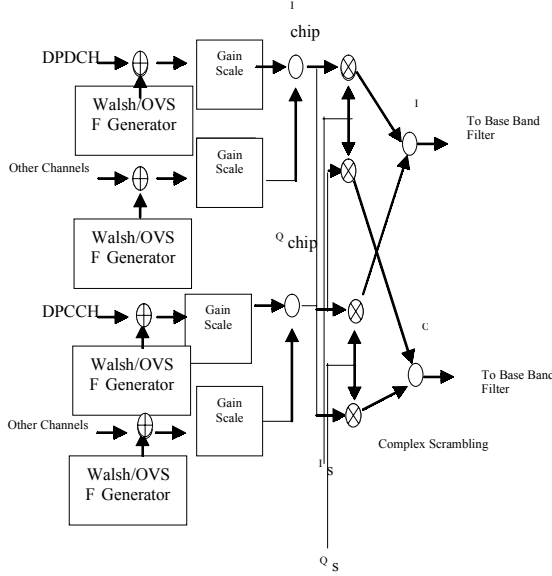
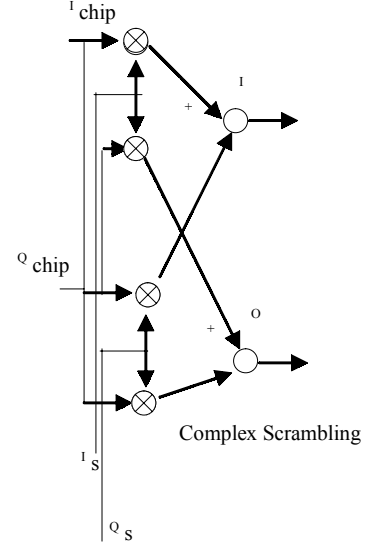


Figure 2. Complex scrambling block diagram



are later I/Q modulated, so they can be expressed as complex signals:

$$\begin{aligned}
 3. \quad I + jQ &= (I_{\text{chip}} \cdot I_s - Q_{\text{chip}} - Q_s) \\
 &\quad + j(I_{\text{chip}} - Q_s + Q_{\text{chip}} \cdot I_s) \\
 &= (I_{\text{chip}} + jQ_{\text{chip}}) \cdot (I_s + jQ_s) \\
 &= A_{\text{chip}} - A_s - e^{j\phi_s}
 \end{aligned}$$

where A_{chip} and $e^{j\phi_{\text{chip}}}$ are the amplitude and the phase of the $I_{\text{chip}} + jQ_{\text{chip}}$ signal; A_s and $e^{j\phi_s}$ are the amplitude and the phase of the $I_s + jQ_s$ signal.

Therefore the amplitude (A) of the resulting $I + jQ$ signal is the product of the amplitudes of both signals. Its phase is the sum of their phases.

This section refers to the original complex data signal spread into chips as the *chip signal*, and to its I and Q components as I_{chip} and Q_{chip} respectively. The resulting complex signal can be referred to as the *final signal*, and its I and Q components simply as I and Q respectively. For simplicity, a signal with only two channels

(one in the I path and the other one in the Q data paths) is used to explain this concept in the I/Q plane (see Figure 3). In the case of two channels with the same amplitude (for I_{chip} and Q_{chip}), the chip signal maps onto a QPSK constellation. The scrambling signal also maps onto a QPSK constellation (since the scrambling I_s and Q_s signals have values of 1 or -1). In the case of two channels with different amplitudes, the chip signal maps onto a rectan-

Figure 3. What is the result of complex scrambling the chip and scrambling signals for these two cases?

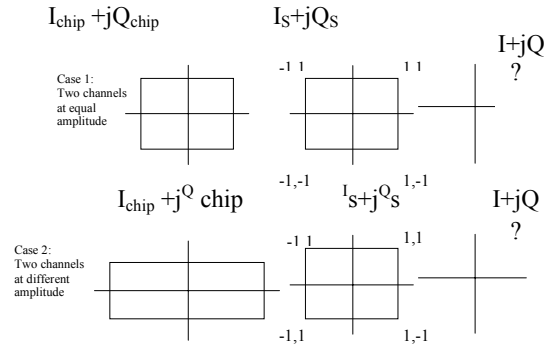
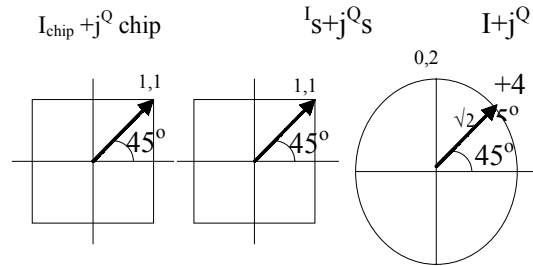


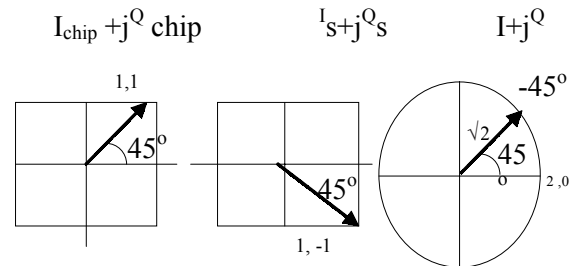
Figure 4a. Scrambling signal at $+45^\circ$



gular 4-QAM constellation. The scrambling signal still corresponds to a QPSK constellation. Figures 4a and 4b illustrate what happens for a single chip point in the original QPSK constellation (Case 1). The original chip signal is the same for both figures (4a and 4b), and the scrambling signal is different. The amplitude of the final signal is the product of the amplitudes of the chip and scrambling signals. The phase of the scrambling signal is added to the phase of the original chip signal (45°). Any point from the original QPSK chip constellation is rotated by 45° , -45° , 135° , or -135° , depending on the values for I_s and Q_s at that time.

Therefore, in the case of two channels with equal amplitudes (QPSK constellation), all the points of the final constellation lie on top of the I or Q axis. The result is a QPSK constellation aligned with the I/Q axes (rotated 45° from the original constellation). In the case of two channels with unequal amplitudes, the amplitude of the resulting constellation is also constant. The chip points from the original 4-QAM constellations are rotated by 45° , -45° , 135° or -135° , since the scrambling signal still corresponds to a QPSK constellation. However, the final constellation has eight points distributed around a circle, since the phases for the chip points in the original constellation are different from 45° , -45° , 135° , and -135° . The angular distribution of the points is determined by the relative amplitudes of the two channels (I_{chip} and Q_{chip}).

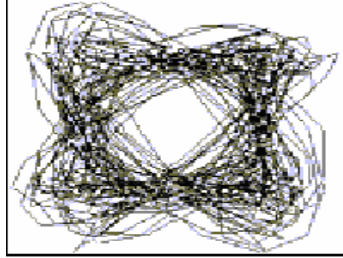
Figure 4b. Scrambling signal at -45°



ORTHOGONAL COMPLEX QUADRATURE PHASE SHIFT KEYING (OCQPSK)

2G systems commonly use modulation formats that limit transitions through zero in the reverse link (uplink). For example, cdmaOne uses OQPSK (Offset Quadrature Phase Shift Keying); this reduces the peak-to-average power ratio of the signal, which allows for a more efficient amplifier, maximizing battery life. However in complex scrambling, if random PN signals are assigned to I_s and Q_s , transitions from any point to any point in the final constellation are possible. This results in a high peak-to-average ratio when compared to existent 2G formats for the reverse link. Figure 5 shows the constellation diagrams of an OQPSK signal and a QPSK signal with basic complex scrambling. The spreading technique that uses basic complex scrambling and PN signals for I_s and Q_s is known as Pseudo-Noise Complex Quadrature Phase Shift Keying (PNCQPSK). 3G systems use Orthogonal Complex Quadrature Phase Shift Keying (OCQPSK) to reduce the peak-to-average power ratio of the signal (Iglesias-Xamani, 2000). OCQPSK is a variation of basic complex scrambling that eliminates zero-crossings for every second chip point. It accomplishes this by using a specific repeating se-

Figure 5a. Constellation of an OQSPK signal



quence (or function) as the scrambling signal and by choosing specific orthogonal codes to spread the different channels (3GPP TS 25.213 V3.6.0, 2001).

HOW DOES OCQPSK WORK?

Basically, OCQPSK uses complex scrambling with a fixed repeating function as the scrambling signal. This function is known as the Walsh rotator and it is defined as $W_0 = \{1, 1\}$ for I_s and $W_1 = \{1, -1\}$ for Q_s . The repeating Walsh rotator sequence ($I_s = W_0 = \{1, 1\}$; $Q_s = W_1 = \{1, -1\}$) is used as the scrambling signal. For two consecutive identical chip points, the first one is rotated by $+45^\circ$ and the second one by -45° , which ensures that they will be 90° apart in the final constellation and the transition between them does not go through zero. This technique assumes pairs of consecutive identical chips. This can be achieved by using only even-numbered Walsh functions to spread the data from the different channels. Even-numbered Walsh functions consist of pairs of identical bits. For example, for a Walsh code length of 8 bits: $W_0 = \{1, 1, 1, 1, 1, 1, 1, 1\}$, $W_2 = \{1, 1, -1, -1, 1, 1, -1, -1\}$, and so forth. Therefore, using only even-numbered Walsh functions to spread the different channels ensures that the chip signal consists of pairs of identical consecutive chips (Iglesias-Xamani, 2000).

Figure 5b. Constellation of a QPSK signal with basic complex scrambling

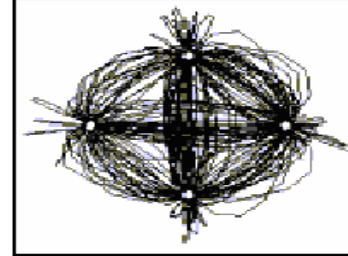


Figure 7 shows a generic block diagram where the Walsh rotator is used as the scrambling signal. To simplify the diagram, only two channels at the same amplitude are transmitted. The orthogonal functions chosen in this case to spread each one of the channels are 8-bit Walsh codes 0 and 2, respectively. For example, if the data signal is $I_D = 1$ and $Q_D = -1$, the chip signal is $I_{chip} = 1, 1, 1, 1, 1, 1, 1, 1$ and $Q_{chip} = -1, -1, 1, 1, -1, -1, 1, 1$. The constellation of this signal consist of the complex points $I_{chip} + jQ_{chip} = \{1 - j1, 1 - j1, 1 + j1, 1 + j1, 1 - j1, 1 - j1, 1 + j1, 1 + j1\}$. Therefore, the chip signal consists of pairs of identical consecutive points. Each pair is multiplied with the scrambling signal formed by the Walsh rotator $I_s + jQ_s = W_0 + jW_1 = \{1 + j1, 1 - j1\}$.

Therefore, the first point in the pair is phase shifted by $+45^\circ$, and the second one by -45° , which ensures that they will be 90° apart in the final constellation (see Figure 6). In the real system, the length of the orthogonal code for each channel depends on that channel's data rate. OCQPSK limits the choice of available orthogonal spreading codes. However, this limitation does not place a large constraint, because a single mobile does not need to support a large number of traffic channels. In most cases, there are sufficient orthogonal codes to handle most channel configurations for both cdma2000 and W-CDMA (Iglesias-Xamani, 2001). In OCQPSK, transitions through zero

Figure 6. Complex rotator first and second chip points by $+45^\circ$ and -45°

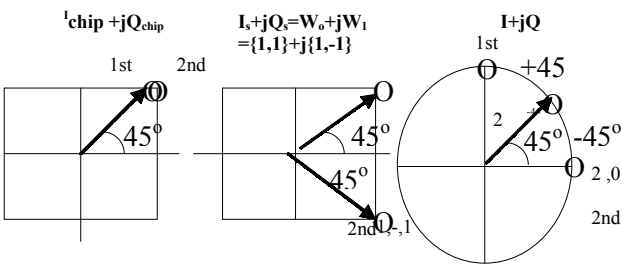


Figure 7. Generic block diagram where the Walsh rotator is used as the scrambling signal

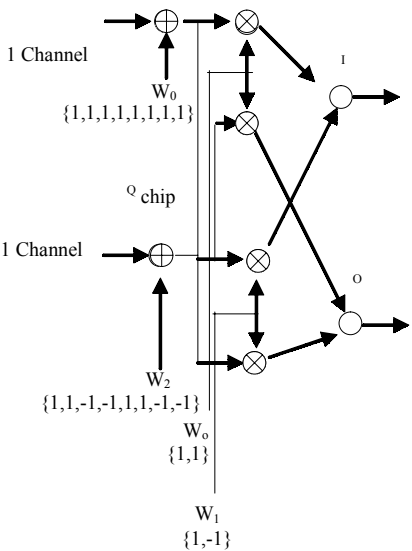


Figure 8a. Constellation of signal without OCQPSK spreading

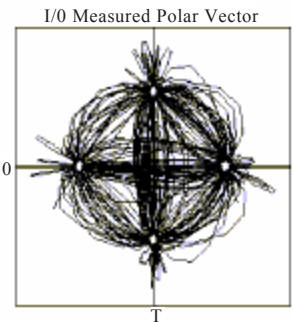


Figure 8b. Constellation of signal with OCQPSK

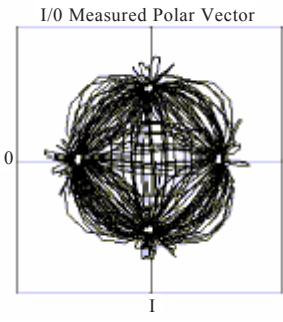
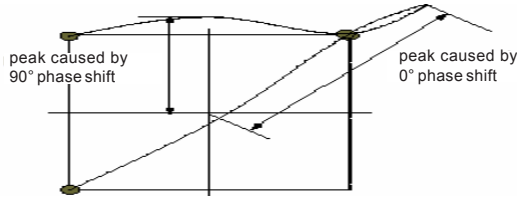


Figure 9. 0° phase shift transition cause higher signal overshoot

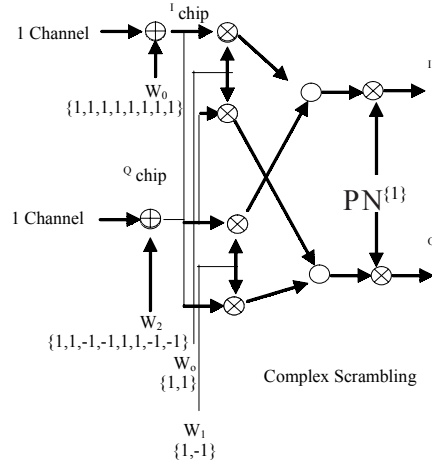


are only eliminated for pairs of consecutive points. Transitions across pairs may go through zero. For example, for four consecutive points, the transition between the first and the second points (or the third and the fourth) does not go through zero. However, the transition between the second and the third point may go through zero. The basic idea behind OCQPSK is to minimize zero-crossings to improve the peak-to-average power ratio of the signal. Figure 8 shows the constellations of two signals: the first one without OCQPSK spreading, and the second one with OCQPSK spreading. In addition to minimizing zero-crossings, OCQPSK eliminates 0° phase shift transitions for every second chip point. A 0° phase transition occurs when two consecutive points are at the same place on the final constellation. This causes overshooting trajectory, which increases the peak-to-average power ratio of the signal as shown in Figure 9. OCQPSK forces 90° transitions between pairs of consecutive points. This minimizes 0° phase transitions, which further reduces the peak-to-average power ratio of the signal.

PRIMARY PN FUNCTION

As shown so far, OCQPSK uses a Walsh rotator and the right choice of orthogonal functions for the different channels to minimize

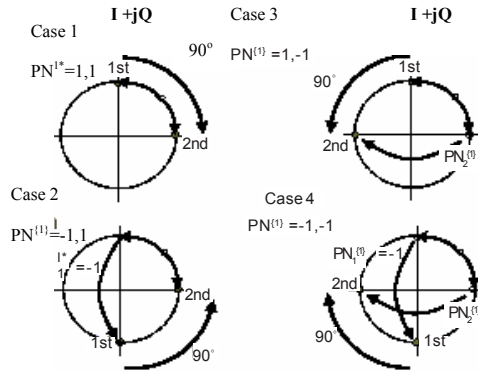
Figure 10. OCQPSK: complex scrambling; Walsh rotator, even-numbered Walsh, and primary PN sequence



zero-crossings and 0° transitions in the final constellation. This improves the peak-to-average power ratio. The real OCQPSK block diagram is more complex. After complex multiplication with the Walsh rotator, a primary PN spreading code $PN^{(1)}$ is applied to the final I and Q signals to allow for identification of the mobile and correlation at the receiver. The $PN^{(1)}$ sequence is the same for I and Q and it does not affect the number of 90° transitions. The $PN^{(1)}$ spreading code does not necessarily have to be applied after the complex scrambling. It can instead be directly multiplied with the I and Q components of the scrambling signal before the complex scrambling. The final result is the same in both cases.

Figure 11 shows why the $PN^{(1)}$ sequence does not affect the number of 90° transitions. The $PN^{(1)}$ sequence is always the same for I and Q. Therefore, a value of $PN^{(1)}$ of +1 does not change the location of the constellation point. A value of -1 inverts the location of the final constellation point. In this example, the final constellation points (after complex scrambling with the Walsh rotator and before apply-

Figure 11. Effect of primary PN sequence



ing the $PN^{(1)}$ sequence) correspond to (0,1) and (1,0). The figure shows the four possible cases.

In the first case, the value of $PN^{(1)}$ is +1 for both points. Therefore the constellation does not change. In the second case, the value of $PN^{(1)}$ is -1 for the first point in a pair and +1 for the second point. The location of the first constellation point changes, but the resulting constellation points are still 90° apart. In the third case, the value of $PN^{(1)}$ is +1 for the first point in a pair and -1 for the second point. Again, the resulting constellation points are 90° apart. In the fourth case, the value of $PN^{(1)}$ is -1 for both the first and second points. The location of both constellation points is inverted, but the

resulting constellation points are still 90° apart.

SECONDARY PN FUNCTION

A decimated secondary PN spreading code (P) is multiplied with the Q component of the Walsh rotator ($W_1 = \{1, -1\}$), as shown in Figure 11. The secondary PN spreading code minimizes Multi-Access Interference (MAI), thereby improving reception at the receiver. P is a decimated version of the real chip rate sequence $PN^{(2)}$. For example, for a decimation factor of two, P holds its value for two chip time periods, which effectively makes its rate half the chip rate. P randomizes the direction of the phase rotation while keeping the phase difference of 90° between pairs of consecutive final points.

Figure 12 illustrates the effect of P in the final constellation. In the first case, $P = 1$. Therefore, the Q component of the scrambling signal is $W_1 = \{1, -1\}$. The first chip point is rotated by +45° and the second by -45°. In the second case, $P = -1$. Therefore, the Q component of the scrambling signal is now $\{-1, 1\}$. The first chip point is rotated by -45° and the second by +45°. This still avoids zero-crossings and 0° transitions between these two chip points.

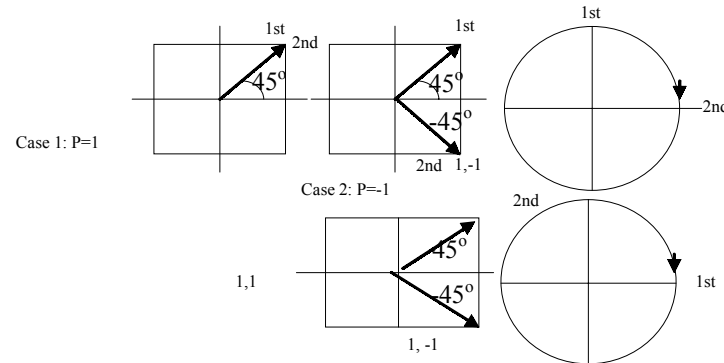
 Figure 12. The decimated secondary PN sequence (P) randomizes the direction of the rotation


Figure 13a. The proposed uplink structure for W-CDMA systems

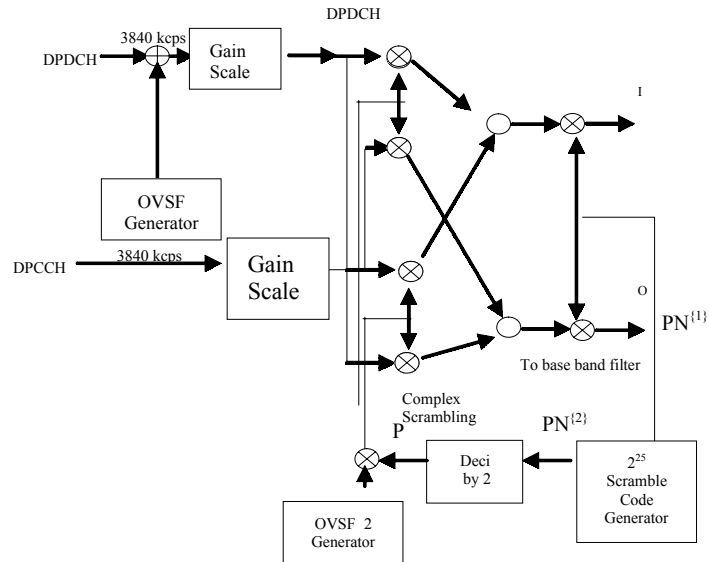
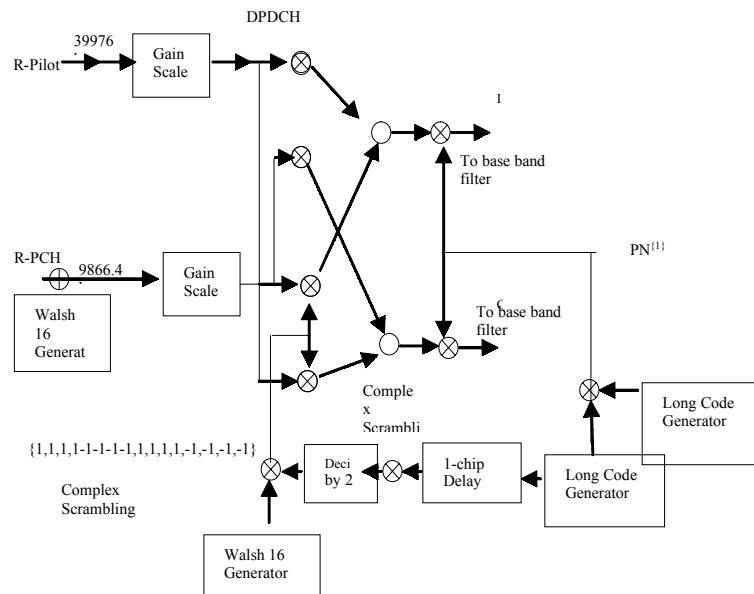


Figure 13b. The proposed uplink structure cdma2000



FINAL OCQPSK SPREADING (SCRAMBLING) FUNCTION

When PN (1) and P are included, the total spreading (or scrambling) function for OCQPSK is:

$$I_s + jQ_s = PN^{(1)} (W_0 + jP W_1)$$

where: $W_0 = \{1,1\}$; and $W_1 = \{1,-1\}$.

Since $W_0 = \{1,1\}$, this function can be removed from the equation. The function can then be expressed as:

$$I_s + jQ_s = PN^{(1)} + jPN^{(1)} P W_1$$

where: $W_0 = \{1,1\}$; and $W_1 = \{1,-1\}$.

Figure 13(a) shows the proposed uplink structure for W-CDMA systems. In W-CDMA, Orthogonal Variable Spreading Functions (OVSFs) are used instead of Walsh codes. The DPCCH is always spread with a 256-bit code 0 ($C_{256,0}$), which corresponds to $\{1,1,1,1,1,1,\dots\}$, so it does not need to be implemented in the block diagram. When a single traffic channel is transmitted, the OVSF code for that channel depends on its data rate. Because OVSF 0 = $W_0 = \{1,1\}$, this section of the Walsh (or OVSF) rotator does not need to be implemented in the block diagram. A 2^{25} gold code generator is used to obtain the two pseudorandom sequences. The generator uses two sets of shift registers to generate the gold codes.

Figure 13(b) shows the block diagram for the reverse link of a cdma2000 Spread Rate 3 (SR3) system with the R-Pilot and a traffic channel. In cdma2000, the Pilot is not spread with any Walsh code, which corresponds to Walsh code 0, $\{1,1,1,1,1,1,\dots\}$. The fundamental traffic channel is always spread with Walsh 16 code 4, $\{1,1,1,1,-1,-1,-1,-1,1,1,1,1,-1,-1,-1,-1\}$. The PN (1) is formed by applying a user's 42-bit code mask to the original cdmaOne

42-bit long code, and scrambling the result against the cdmaOne I short code. P is created by delaying the long code by one chip, scrambling it with original cdmaOne Q short code, and decimating by a factor of two. The names PN (1), PN (2), and P are not used in W-CDMA or cdma2000, but have been added to both block diagrams for clarification.

SIMULATION RESULTS

The following scopes display the signal in various ways. Time Scopes show the bit stream before spreading, after spreading, and after combining the different weighted physical channels. It shows both the real and the imaginary part separately. It also displays both the real and the imaginary part of the output of the channel estimator. Figure 14 shows CCDF comparison between signals with and without OCQPSK $\{1, -1\}$. The probability of a zero-crossing for a regular QPSK or PNCQPSK signal with two channels at the same amplitude is 1/4. For OCQPSK, the probability of a zero crossing is limited to every other chip point, and is therefore reduced in half (1/8). The probability of 0° phase shift transitions is also reduced from 1/4 to 1/8. All this improves the peak-to-average power ratio of the signal by approximately 1 to 1.5 dB. OCQPSK spreading is still advantageous when multiple channels at different amplitudes are transmitted. Figure 14 shows the CCDF (Complementary Cumulative Distribution Function) for two signals.

A Signal with Basic Complex Scrambling

This signal has been generated from cdma2000 SR1 forward link signal with a single traffic channel (which uses basic complex scram-

Orthogonal Complex Quadrature Phase Shift Keying (OCQPSK) Spreading for 3G W-CDMA Systems

Figure 14. CCDF comparison between signals with and without OCQPSK $\{1, -1\}$

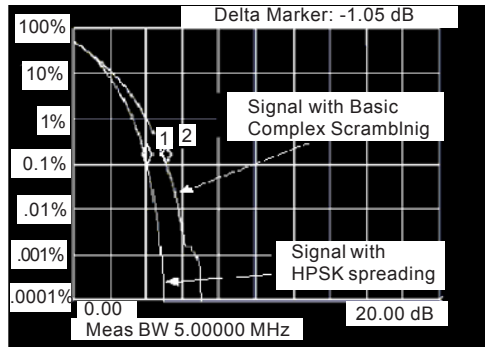


Figure 15. Bits before and after spreading (real part)

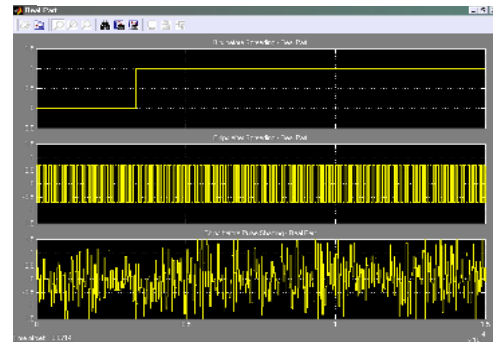


Figure 16. Bits before and after spreading (imaginary part)

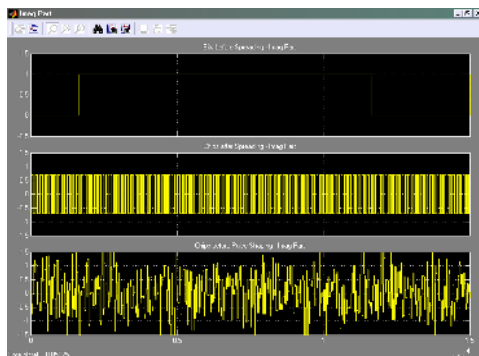


Figure 17. Power spectrum of the signal before spreading

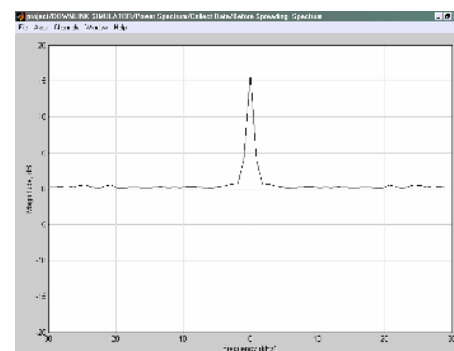


Figure 18. Power spectrum of the signal after spreading

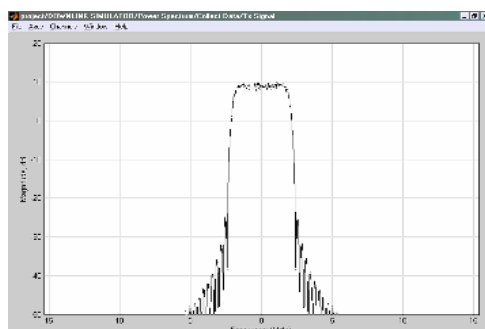
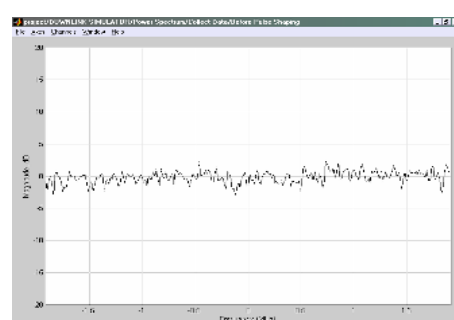


Figure 19. Power spectrum of the signal after pulse shaping



bling). The filtering has been modified to cdma2000 reverse link filtering (which, as opposed to the forward link, does not include an equalization function).

A Signal with OCQPSK Spreading

This signal is a cdma2000 SR1 reverse link signal with two channels (R-Pilot and R-FCH) at the same power level. The CCDF curve provides the distribution of particular peak-to-average ratios versus probability. In this case, for a probability of 0.1%, the peak-to-average ratio of the OCQPSK signal is about 1 dB lower than the signal with basic complex scrambling; this plot does not provide a straight comparison between basic complex scrambling and OCQPSK. The coding between forward and reverse link signals is different, which may impact the results. However, the plot is indicative of the performance of OCQPSK spreading versus basic complex scrambling. Signals with high peak-to-average power ratio may saturate the power amplifier, causing higher interference in the adjacent channels and a reduction of system capacity. To minimize this, the amplifier must be designed with a larger back off, which in turn reduces amplifier efficiency. Therefore, high peak-to-average power ratios reduce battery life, one of the critical characteristics of the mobile phone. Figures 15 and 16 show bits before and after spreading (real and imaginary part). Figures 17 and 18 show the power spectrum of the signal before and after spreading. Figure 19 shows the power spectrum of the signal after pulse shaping.

CONCLUSION AND FUTURE WORK

OCQPSK spreading technique in a 3G W-CDMA system is studied in this chapter. We

proposed a reverse link and uplink structure for the third-generation W-CDMA system. The proposed structure is based on OCQPSK spreading technique. We show some results of the simulator. A convenient or optimal selection of the parameters according to the mobile network environment (mobiles speeds characteristics, other statistics) is left for future research

This chapter also explained why OCQPSK has been selected as the spreading technique for the reverse link (uplink) of 3G systems such as W-CDMA and cdma2000. The main points are:

- In W-CDMA and cdma2000 systems, the mobile phone can transmit multiple I/Q multiplexed channels at different power levels.
- Complex scrambling facilitates this by distributing the power evenly between the axes.
- OCQPSK is a variation of complex scrambling that uses a Walsh rotator and specific orthogonal (Walsh or OVSF) spreading functions to minimize zero-crossings and 0° phase shift transitions. This improves the peak-to-average power ratio of the signal.
- The chapter shows how to start making modulation quality measurements on the reverse link of W-CDMA and cdma2000 signals using existing instrumentation.

Furthermore we simulated a complete downlink physical layer model showing various results of BER and BLER calculation, and also various time scopes and power spectrums. The simulator was designed to satisfy most of the parameters mentioned in the 3GPP specifications. The simulator was tested against various variable parameter values of measurement chan-

nels and propagation conditions, as 3GPP standard specifies minimum requirement tests for different data rates under different propagation conditions.

Areas for Future Work

Third-generation mobile systems based on W-CDMA are currently being rolled out all over the world. The services offered by these systems can be divided into two groups, real-time and non-real-time, depending on the requirements on transfer delay posed by the applications. Examples of real-time services are speech, videoconferencing, and streaming video clips. Messaging and Web browsing are two examples of non-real-time services. Initially these networks provide real-time services of up to 64 kbps and non-real-time services of up to 384 kbps.

In parallel with the rollout of the first 3G networks, the evolution of W-CDMA is in full progress. In Release 5 of the W-CDMA specifications, a new downlink channel called high-speed downlink shared channel is introduced. The new channel provides peak bit rates in the order of 10 Mbps as well as increased system capacity. The primary target services in the development of the new channel have been non-real-time services, although recently different mechanisms for supporting real-time services have been proposed.

The strict delay requirements associated with real-time services, like streaming and speech, can only be fulfilled if radio resource management functions such as admission control and scheduling functions distribute the scarce radio resources among the different connections in a controlled manner. In addition, the resource handling must be efficient in order to serve many users. Today, such algorithms do not exist for the high-speed downlink shared channel.

A number of scenarios can be investigated where different kinds of strategies and algorithms are employed to efficiently fulfill the real-time requirements. The simplest scenario to be studied is one where all real-time users require the same streaming service, and the streaming data is sent over dedicated channels. This is the easiest scenario from a quality of service perspective, and could be regarded as a reference case. This scenario should be compared with one where also real-time data is sent over the high-speed downlink shared channel. Other interesting scenarios to be investigated are where different streaming users request different bit rates (e.g., 56 kbps or 300 kbps). In all scenarios, any non-real-time data is sent over the high-speed downlink shared channel.

REFERENCES

- 3G TS 25.101 V3.7.0. (2001, June). *UE radio transmission and reception (FDD)*.
- 3G TS 25.104 V3.7.0. (2001, June). *UTRA (BS) FDD; radio transmission and reception*.
- 3G TS 25.211 V3.7.0. (2001, June). *Physical channels and mapping of transport channels onto physical channels (FDD)*.
- 3G TS 25.213 V3.6.0. (2001, June). *Spreading and modulation (FDD)*.
- 3GPP TS 25.101 V5.2.0. (2002-2003). *UE radio transmission & reception (FDD)*. Third-Generation Partnership Project Technical Specification Group Radio Access Network Working Group 1.
- 3GPP TS 25.201 V5.0.0. (2001-2002). *Physical layer—general description*. Third-Generation Partnership Project Technical Specification Group Radio Access Network Working Group 1.

- 3GPP TS 25.211 V5.0.0. (2002-2003). *Physical channels and mapping of transport channels onto physical channels (FDD)*. Third-Generation Partnership Project Technical Specification Group Radio Access Network Working Group 1.
- 3GPP TS 25.212 V5.0.0. (2002-2003). *Multiplexing and channel coding (FDD)*. Third-Generation Partnership Project Technical Specification Group Radio Access Network Working Group 1.
- 3GPP TS 25.213 V2.3.0. (1999). *Spreading and modulation*. Third-Generation Partnership Project Technical Specification Group Radio Access Network Working Group 1.
- 3GPP TS 25.214 V5.0.0. (2002-2003). *Physical Layer Procedures (FDD)*. Third-Generation Partnership Project Technical Specification Group Radio Access Network Working Group 1.
- Dahlman, E., Gudmundson, J., Nilsson, M., & Skold, J. (1998). UMTS/IMT-2000 based on wideband CDMA. *IEEE Communications Magazine*, 36(September), 70-80.
- Dinan, E. H., & Jabbari, B. (1998). Spreading codes for direct sequence CDMA and wideband CDMA cellular networks. *IEEE Communications Magazine*, 36(September), 48-54.
- Holma, H., & Toskala, A. (2000). *W-CDMA for UMTS: Radio access for third-generation mobile communications*. New York: John Wiley & Sons.
- Iglesias-Xamani, M. (2000). Hybrid-Phase-Shift Keying proposed for 3G systems. *Wireless Systems Design*, 5(1), 24-32.
- Laird, K., Whinnett, N., & Buljore, S. (1999). A peak-to-average power reduction method for third-generation CDMA reverse links. *Proceedings of the IEEE 49th Vehicular Technology Conference* (Vol. 1, pp. 551 -555).
- Mathworks. (n.d.). Retrieved from www.mathworks.com
- Ojanpera, T., & Prasad, R. (1998). An overview of air interface multiple access for IMT-2000/UMTS. *IEEE Communications Magazine*, 36(September), 88-95.
- Proakis, J. G. (1995). *Digital communications* (3rd ed.). New York: McGraw-Hill.
- Xu, B., Vu, T. B., & Mehrpour, H. (1999). A space-time rake receiver for synchronous DS-CDMA system based on smart antenna. *Proceedings of the IEEE Vehicular Technology Conference*.

Chapter XIII

Code and Time Synchronization of the Cell Search Design Influence on W-CDMA System Performance

Shailendra Mishra

Dehradun Institute of Technology, India

ABSTRACT

The large-scale statistics of an improved cell search design (improved CSD) using cyclic codes is compared with the 3GPP cell search design using comma free codes (3GPP-comma free CSD) in terms of acquisition time for different probabilities of false alarm rates and to achieve faster synchronization at lower hardware complexity is addressed in this chapter. In the chapter we also proposes design improvements in stage 2 of the 3GPP-comma free CSD. The 3GPP-comma free CSD proposed in this chapter uses a Fast Hadamard Transformer (FHT) in stage 2 that achieves lower hardware complexity and faster decoding. Furthermore, masking functions are used in stage 3 of both the improved CSD and the 3GPP-comma free CSD to reduce the number of scrambling code generators required as described in previous work. This results in a reduction in the ROM size required to store the initial phases of the scrambling code generators in stage 3. The Improved CSD proposed in this chapter aims to achieve faster synchronization between the mobile station (MS) and the base station (BS) and thus improves system performance. Our results indicate that for a channel whose signal-to-noise ratio is degraded with additive white Gaussian noise (AWGN), the improved CSD achieves faster synchronization with the base station and has lower hardware utilization when compared with the 3GPP-comma free CSD scheme under the same design constraints.

INTRODUCTION

W-CDMA is one of the leading wideband digital cellular technologies that will be used for the third-generation (3G) cellular market. The earlier Japanese W-CDMA trial system and the European Universal Mobile Telephone System (UMTS) have both served as a foundation for the workings of this harmonized W-CDMA system under the supervision of the Third-Generation Partnership Project (3GPP). The 3GPP organizational partners are the European Telecommunications Standard Institute (ETSI), the Japanese Association of Radio Industries and Businesses (ARIB), the Japanese Telecommunication Technology Committee (TTC), the Korean Telecommunications Technology Association (TTA), and the American Standards Committee on T1 Telecommunications. The harmonized system is sometimes referred to as 3GPP W-CDMA, to distinguish it from earlier wideband CDMA versions. The W-CDMA system will employ wideband CDMA in both frequency division duplex (FDD) and time division duplex (TDD) modes (Ojanperä & Prasad, 1998).

The main difference between W-CDMA and CDMA2000 is that W-CDMA supports asynchronous base stations (BSs) whereas CDMA2000 relies on synchronized BSs. Synchronous CDMA systems need an external

time reference. A global positioning system (GPS) clock can be used by all BSs to synchronize their operations. This allows the mobile station (MS) to use different phases of the same scrambling code to distinguish between adjacent BSs. In an *asynchronous* CDMA system, each BS has an independent time reference, and the MS does not have prior knowledge of the relative time difference between various BSs. The advantage of asynchronous operation is that it eliminates the need to synchronize the BSs to an accurate external timing source. However, since there is no external time synchronization between the adjacent BSs, different phases of the same code cannot be used to distinguish adjacent BSs. Thus, in an *asynchronous* CDMA system, adjacent BSs can only be identified by using distinct scrambling codes. Consequently, cell search—which involves the process of achieving code, time, and frequency synchronization of the MS with the BS—takes longer in comparison to a *synchronous* CDMA system. Cell search is complicated in the presence of signals which are intended for other mobile systems within a cell, as well as signals from other BSs (Dahlman et al., 1998). Thus, it is very important to develop algorithms and hardware implementations to perform cell search using lower acquisition time and minimum hardware resources for *asynchronous* CDMA systems. In W-CDMA,

Figure 1(a). Downlink channel multiplexing and spreading

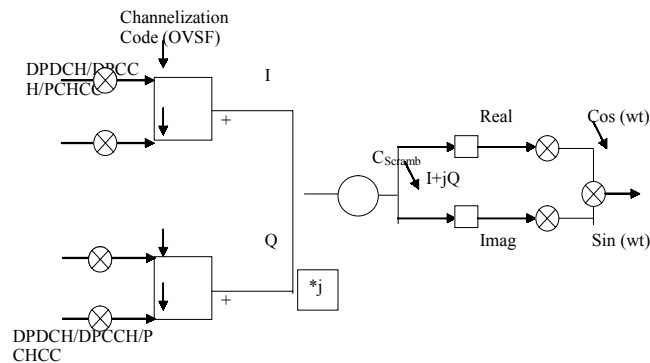
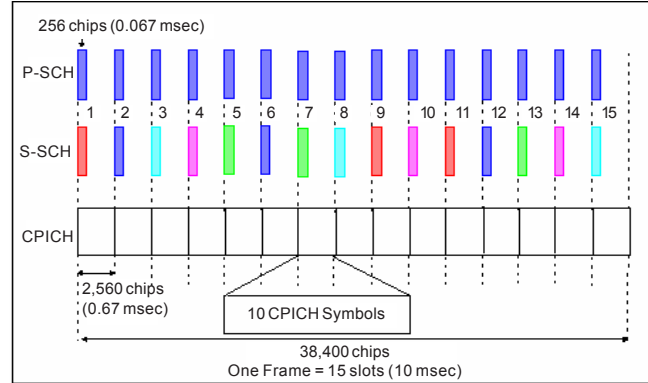


Figure 1(b). Frame/slot structures for CPICH, P-SCH, and S-SCH



a cell is identified mainly by its downlink scrambling code as shown in Figure 1(a).

There are 512 primary downlink scrambling codes reused throughout a system. These 512 codes are based on length $2^{18}-1$ Gold sequence truncated to one frame interval which is 38,400 chips for the chip rate 3.84. To reduce the complexity of searching through the 512 downlink primary scrambling codes, the concept of code grouping and use of code group indicator code (GIC) were introduced in Higuchi, Sawahashi, and Adachi (1997). The scrambling code is identified by first identifying its code group to significantly reduce the degree of code uncertainty. The complexity of cell search is further reduced by combining code group identification and frame boundary synchronization into one stage (Nystrom, Jamal, Wang, & Esmailzadeh, 1998). With this scheme, the time uncertainty is completely resolved when the code group identity is obtained. As a result, the complexity of identifying the scrambling code is identified code group is significantly reduced schemes with further complexity reduction by increasing the number of code groups proposed in Nystrom et al. (1998). To facilitate cell search, three channels are used—the primary synchronization channel (P-SCH), the secondary synchronization channel (S-SCH), and the

common pilot channel (CPICH). Each frame of 38,400 chips (or 10ms) is separated into 15 slots, each with 25,600 chips (or 0.67ms). The CPICH, which used to carry the downlink common pilot symbols, is scrambled by the primary downlink scrambling code of the cell. Within each CPICH timeslot, there are 10 pilot symbols, each spread by 256 chips. All symbols are Quadrature Phase Shift Keying (QPSK) modulated, and the modulation values of the pilot symbols are known once the mobile system knows the frame boundary. The spreading sequence of CPICH is taken from the set of orthogonal variable spreading factor (OVSF) codes, maintaining mutual orthogonality between CPICH and other downlink channels also spreading OVSF codes. Unlike CPICH, neither the P-SCH nor the S-SCH is scrambled by the primary downlink scrambling code. Instead of the OVSF codes, another sequence of length 256 chips are used. The P-SCH sequence is transmitted once in the same position in every slot, and can thus be used for detecting the slot boundary. Frame/slot structures for CPICH, P-SCH, and S-SCH are shown in Figure 1(b).

Furthermore, all cells use the same P-SCH matched filter to detect the slot boundaries of downlink signals.

Cell search is performed according to the algorithm proposed by Wang and Ottosson (2000). In the proposed cell search algorithm, code and time synchronization is achieved assuming a large frequency error, and after achieving code and time synchronization, frequency synchronization is performed. In this study we consider the problem of achieving code and time synchronization. The process of achieving code and time synchronization in the cell search algorithm for W-CDMA systems is divided into three stages: (1) slot synchronization, (2) frame synchronization and code group identification, and (3) scrambling code identification.

CELL SEARCH DESIGN

Cell search design is critical as it impacts the system performance and there is a need to design efficient receiver structures and algorithms to reduce the cell search time. This section summarizes efforts by research groups and the 3GPP working groups to design efficient schemes and algorithms for each of the three stages of the cell search algorithm. Wang and Ottosson (2000) proposed a pipelined process to be used in the first three stages of the cell search algorithm. The cell search scenarios considered in their study are: (1) initial cell search, when a mobile is switched on; and (2) target cell search, during idle and active modes of the MS. Instead of the serial cell search sequentially searching through code, time, and frequency, their method first acquires code and time synchronization assuming a larger frequency error and then performs frequency synchronization (Peterson, Ziemer, & Borth, 1995; Wang & Ottosson, 2000). The synchronization code sequences used in stages 1 and 2 of the cell search algorithm are made up of bits called “chips,” which can be either +1 or -1. The synchronization code sequences are 256

chips in length. If a traditional matched filter is used, then a huge adder circuit (256 input adder) will be required to sum up the correlation results. This will lead to wastage of hardware resources. Hence, Siemens and Texas Instruments in their working group draft have suggested a hierarchical matched filter design, which uses two matched filters to reduce the hardware complexity significantly. The 3GPP specification uses comma free codes in stage 2 of the cell search algorithm (3GPP RAN TS 25.213 v4.0.0, 2001-2003). Nortel Networks, in their working group proposal, suggested the use of cyclic codes in the SCH (synchronization channel) (3GPP RAN TS 25.214 v4.0.0, 2001-2003). The cyclic codes can reduce hardware utilization and acquisition time if the receiver is properly designed. To reduce the complexity of searching through all of the 512 scrambling codes, the concept of code grouping and group indicator codes (GICs) was introduced (Higuchi et al., 1997). This reduces the cell search time, as the scrambling code is identified by first detecting the code group. Once the code group is detected, the scrambling code used by the cell can be easily identified, as there are a limited number of codes in each code group. This reduces the cell search time significantly. This idea is accepted in the 3GPP specifications. To further reduce cell search time, frame boundary synchronization is also achieved in stage 2 after identifying the code group and slot ID (Nystrom et al., 1998). Ericsson, in its working group draft, proposed increasing the number of code groups in stage 2 of the cell search (TSGR1-6/99). Increasing the number of code groups reduces the number of scrambling codes in a code group. The proposed scheme uses either 256, 128, or 64 code groups in stage 2 of the cell search; the group claims that the scheme using 256 code groups is the preferred scheme as it requires only two scrambling code correlators in stage 3 of initial cell

search and achieves reduced hardware complexity.

In stage 2 of the 3GPP-comma free CSD, a FHT design is proposed to replace the Golay correlator presented by Li, Sheen, Ho, and Chu (2002). An FHT provides an efficient technique to detect the code group and slot ID in stage 2. Previous FHT designs (Amira, Bouridane, Milligan, & Roula, 2001; Nayak & Meher, 1999) utilize many hardware resources, hence a fast and efficient Hadamard transformer is needed to reduce the hardware utilization and to perform faster decoding.

A compact and efficient FHT design will also draw less power from the handset. Siemens, in its working group draft, has suggested the use of masking functions in stage 3 to reduce the design complexity for generating the scrambling codes in parallel (TSGR1-7/99). The use of masking functions reduces the number of scrambling code generators required to generate the codes in parallel. Any masking function can be selected by the designer as long as they generate codes with minimum overlap. The use of masking functions reduces the hardware significantly as compared to the previous design by. Li et al. (2001) designed an application-specific integrated circuit (ASIC) for performing cell search in W-CDMA systems. In stages 1 and 2 of their cell search design, the authors use a correlator structure to detect the code group and slot ID. The correlator structure used is a Golay correlator (Popovic, 1999). In stage 3 of the cell search algorithm, 16 scrambling code generators are used for generating the codes in parallel.

Cell Search Algorithm

The process of achieving code and time synchronization in the cell search algorithm is divided into three stages: (1) slot synchronization, (2) frame synchronization and code group

identification, and (3) scrambling code identification.

Stage 1: Slot Synchronization

During stage 1 of the cell search procedure, the MS uses the SCHs Primary Synchronization Code (PSC) to acquire slot synchronization to a cell. This is typically done with a single matched filter matched to the PSC, which is common to all cells. The slot timing of the cell can be obtained by detecting peak values in the matched filter output. The starting position of the synchronization code may be determined from observations over one slot duration. However, decisions based on observations over a single slot may be unreliable, when the signal-to-noise ratio (SNR) is low or if fading is severe. Reliable slot synchronization is required to minimize cell search time. In order to increase reliability, observations are made over multiple slots and the results are then combined. This ensures that the correct slot boundary is identified.

Stage 2: Frame Synchronization and Code Group Identification

During stage 2 of the cell search procedure, the MS uses the SCHs Secondary Synchronization Code (SSC) to achieve frame synchronization and identify the code group of the cell found in stage 1. This is done by correlating the received signal with all possible SSC sequences and identifying the maximum correlation value. Since the cyclic shifts of the sequences are unique, the code group as well as the frame synchronization is determined.

Stage 3: Scrambling Code Identification

During stage 3 of the cell search procedure, the MS determines the exact primary scrambling

code used by the cell. The primary scrambling code is typically identified through symbol-by-symbol correlation over the CPICH with all codes within the code group identified in stage 2. In this stage, a threshold value is used to decide whether the code has been identified. The threshold value can be predetermined using a parameter called probability of false alarm rate. This three-stage cell search algorithm helps in simplifying the synchronization process of the MS with the BS.

IMPROVED CELL SEARCH DESIGN

The Improved CSD uses a set of cyclic codes; these were proposed by Nortel Networks to be used on the Secondary SCH. These cyclic codes allow very efficient detection and improve the cell search in terms of acquisition time and hardware utilization.

Stage 1: Slot Synchronization

The MS first needs to acquire the PSC, which is common to all the BSs. These codes are of length 256 chips. The matched filter output is given by,

$$Y = \sum_{j=0}^{255} R_j C_{p_j}$$

where R_j is the j th sample of the received complex signal, and C_{p_j} is the j th bit of the PSC (primary synchronization code).

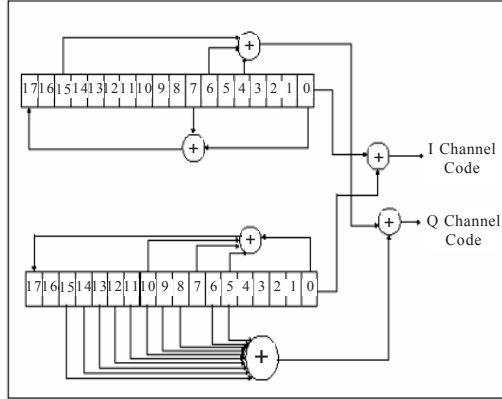
Hence a traditional matched filter implementation would require 256 taps and a large adder circuit. This would increase the delay as well as power consumption at the receiver, which is not desirable. Thus, a hierarchical structure is proposed for performing the matched filter operations, which will need a lesser number of taps, reduced circuitry, and lower power

consumption. The PSC consists of an unmodulated hierarchical sequence of length 256 chips transmitted once every slot. The PSC is the same for every BS in the system and is transmitted time aligned with the slot boundary. The PSC is chosen to have good auto-correlation properties. This means that when the PSC sequence is correlated with itself, the interference from adjacent BSs is minimized and a high peak value is obtained. The hierarchical matched filter consists of two concatenated matched filter blocks. This solution is not ideal for two reasons: (1) the matched filter design requires 64 taps, and (2) the design needs a 64-input adder. Hence in stage 1 of both the Improved CSD and the 3GPP-comma free CSD, the hierarchical matched filter using 16 chip and 16-symbol accumulation is used. There are two such hierarchical matched filters for the I and Q channels of the received complex signal, the correlation results over I and Q channels are combined non-coherently over 1 slot duration, and the result is stored in an accumulator which is implemented as a shift register. The output of the accumulator is given to a comparator block to detect the peak value corresponding to the slot boundary of the closest BS, and the MS needs to synchronize with this BS. As the code can be affected by AWGN and fading, accumulation over multiple slots is needed to correctly identify the slot boundary. It is important that the slot boundary is correctly identified in order to avoid the cost of increased acquisition time in case the wrong slot boundary is given to stage 2.

Stage 2: Frame Synchronization and Code Group Identification

The Secondary SCH consists of 15 sequences belonging to a family of cyclic codes (SSCs), each of length 256 chips. These SSCs are transmitted repeatedly in parallel with the Pri-

Figure 2. Scrambling and code generator



mary SCH. The procedure for constructing the cyclic codes is similar to that of the hierarchical sequence for the Primary SCH, except that it uses specific sequences of length 16 from each code group. For the 32 codes groups and 15 slots (in one frame), 512 different cyclic codes with a length of 256 chips each are constructed. In other words, each of the 32 code groups has 16 cyclic codes. This set of 512 (32×16) cyclic codes has good correlation properties that make it good a candidate for the SSCs. Many pairs of cyclic codes are fully orthogonal as the cross correlation is zero; some pairs have small cross-correlation properties. The cross correlation of each cyclic hierarchical sequence $Cs_{i,k}$ with C_p code of Primary SCH is small. These 512 cyclic codes are unique for each code group/slot locations pair. Thus, it is possible to uniquely determine both the scrambling code group and the frame timing in the second stage of the initial cell search.

By identifying the code group/slot location pair that gives the maximum correlation value, the code group as well as the frame synchronization is determined. The output from the matched filter is given to a non-coherent block, which computes the energy over I and Q channels and then gives the result to the comparator module. One slot search period time (2,560 chips) is enough to uniquely identify the correct

code group and the frame timing in the second stage of acquisition, when the signal-to-noise ratio is high. This is one major difference with the 3GPP-comma free CSD, where at least three slots are necessary to uniquely identify the correct code group and frame timing. The Improved CSD also uses a smaller size ROM 32×16 to store the cyclic codes as compared to the 3GPP-comma free CSD which uses a ROM of size 32×60 to store the comma free codes.

Stage 3: Scrambling Code Identification

After achieving code group and frame synchronization, the scrambling code is identified by correlating the symbols in the CPICH with all possible scrambling codes in the code group. The codes are generated using a scrambling code generator and the descrambling operation is carried out using a descrambler.

The scrambling code generator used to generate the long codes is shown in Figure 2. A total of $2^{18} - 1 = 262,143$ scrambling codes, numbered $0, 1, \dots, 262$, can be generated using the code generator. However, not all of the scrambling codes are used. The scrambling codes are divided into 512 sets, each of a primary scrambling code and 15 secondary scrambling codes. The primary scrambling codes consist of scrambling codes $n = 16 \cdot i$ where $i = 0, 1, \dots, 511$. The i^{th} set of secondary scrambling codes consists of scrambling codes $16 \cdot i + k$, where $k = 1, 2, \dots, 15$. There is a one-to-one mapping between each primary scrambling code and 15 secondary scrambling codes in a set such that the i^{th} primary scrambling code corresponds to the i^{th} set of secondary scrambling codes (TSGR1-6/99). The set of primary scrambling codes is further divided into 32 scrambling code groups, each consisting of 16 primary scrambling codes. The j^{th} scrambling code group consists of pri-

many scrambling codes $16 \cdot 16 \cdot j + 16 \cdot k$, where $j=0,1,\dots,31$ and $k=0,1,\dots,14$.

3GPP-COMMA FREE CELL SEARCH DESIGN

Stage 2 of the 3GPP cell search design uses comma free codes. Stages 1 and 3 for the 3GPP-comma free CSD design were kept the same as the Improved CSD to compare stage 2 of both the designs. A Fast Hadamard Transformer (FHT) is proposed to be used in stage 2 of the cell search algorithm. To reduce the hardware utilization of the FHT design, reduced-length Walsh sequences are proposed.

Stage 2 of 3GPP-Comma Free Cell Search Design

In CDMA systems, the BS identifies each user in a cell by a unique scrambling code. In order to minimize the interference in a cell when two users transmit at the same time, orthogonal (Walsh) codes are used. The Walsh codes are generated using a Walsh-Hadamard function. When these Walsh codes are transmitted by the BS, they are affected by interference, fading, and noise, which may be AWGN. At the receiver, a decoding logic is required to correctly determine which of the Walsh codes was the most likely to have been sent. An FHT can be used to provide such a decoding circuitry (Amira et al., 2001).

For comparison with the Improved CSD scheme, which uses 32 code groups, only 32 of the possible 64 code groups are used. The 32 secondary SCH sequences are constructed such that their cyclic shifts are unique—that is, a non-zero cyclic shift less than 15 of any of the 32 sequences is not equivalent to some cyclic shift of any other of the 32 sequences. Also, a non-zero cyclic shift less than 15 of any of the

sequences is not equivalent to itself with any other cyclic shift less than 15. There are a number of stages in the FHT design depending on the length of the Walsh sequence. Each subsequent stage receives an input from the previous stage in half the number of clock cycles required for the previous stage. This is achieved by reducing the length of shift register by a factor of two for each subsequent stage of the FHT. A counter is used as a clock to determine the time interval at which each successive pair of input signals is received by the FHT. The upper shift registers in each of the stages are always enabled, whereas the lower shift registers are enabled by the bits of the counter.

The length of the counter register is dependent on how many stages there are in the FHT. The counter bit C0 is the LSB, and C2 is the MSB. Counter bit C2 is alternately high for four clock cycles and then goes low for four clock cycles (000...011, 100...111). The bit C0 is alternately high and low for each clock cycle (000,001,...etc.). The number of bits in the counter depend on the number of stages, which in turn depends on the length of Walsh-

Figure 3. Comparison of Improved CSD and 3GPP-comma free CSD $P_{FA} = 10^{-3}$

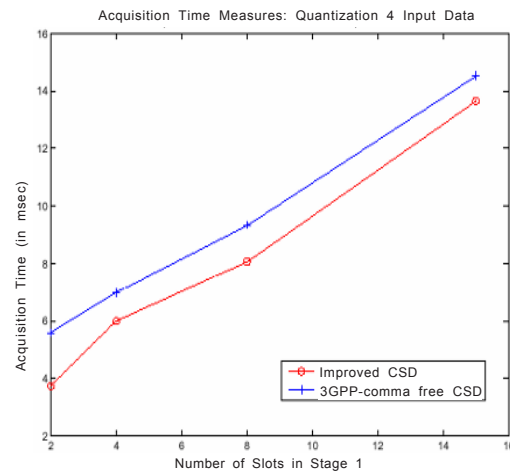
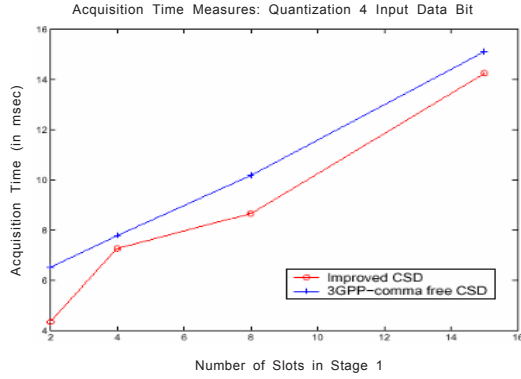


Figure 4. Comparison of Improved CSD and 3GPP-comma free CSD $P_{FA}=10^{-4}$



Hadamard sequence to be used. If there are N Walsh chips, then the counter length must be $\log_2 N$ bits. The length of the shift register in each of the stages of the design is given by the following relation $(N/4)/2^s$. For example, the length of the shift registers used in the first stage of the FHT is $(16/4)/2^0 = 4$. Similarly, the length of registers used in other stages can be calculated.

SIMULATION RESULTS

In this section we measure the acquisition time for both of the cell search designs, Improved CSD, and the 3GPP-comma free CSD. The acquisition time was measured by counting the number of clock cycles used by the RTL simulation. The input chip rate is given by the 3GPP specifications, and this gives the acquisition time measure. To compare the acquisition time between the Improved CSD and the 3GPP-comma free CSD, experiments were carried out using input vectors generated in Matlab. Threshold values determined for the two probabilities of false alarm rates ($P_{FA}=10^{-3}$ and $P_{FA}=10^{-4}$) were 28 and 37 respectively. The number of clock cycles between the start of the

system and the point when the counter in stage 3 exceeds the computed threshold values was determined. The equivalent gate count and maximum frequency of operation were compared for both the designs using a 256-chip sequence in stage 2; it was observed that the Improved CSD uses a fewer number of slots to achieve synchronization as compared to the 3GPP-comma free CSD in stage 2. The results obtained indicate that when averaging is carried out over 15 slots in stage 1 of both the designs ($P_{FA1}=10^{-3}$ and $V_{TH1}=28$), the Improved CSD has an acquisition time of 13.66 msec, as compared to 14.53 msec for the 3GPP-comma free CSD. Thus, the Improved CSD achieves an improvement of 0.87 msec for an AWGN channel (see Figure 3). Similarly, an improvement of 0.87 msec was observed when $P_{FA2}=10^{-4}$ and $V_{TH2}=37$. Figures 4 and 3 show the acquisition time measures for 2, 4, 8, and 15 slots in stage 1 of the design. The number of slots in the other stages were kept fixed as one slot in stage 2 of the Improved CSD, three slots in 3GPP-comma free CSD, and 15 slots in stage 3.

CONCLUSION AND FUTURE WORK

For an AWGN channel model in a high signal-to-noise ratio environment, it was found that accumulation over one slot in the Improved CSD scheme and accumulation over three slots in the 3GPP-comma free CSD scheme in stage 2 of the cell search algorithm give correct code group and slot boundary identification. Due to the reduction in the required number of slots, the Improved CSD uses a lesser number of clock cycles in stage 2, as compared to the 3GPP-comma free CSD to detect the code group and slot ID. This reduction in the number of clock cycles leads to faster acquisition, fewer calls getting dropped, and lower power consumption during the synchronization be-

tween the MS and the BS. The use of cyclic codes in the Improved CSD has lower hardware utilization and a higher maximum frequency of operation, as compared to the 3GPP-comma free CSD. In conclusion, the Improved CSD is a better cell search design in comparison to the 3GPP-comma free CSD since it has faster acquisition time and lower hardware utilization.

This section investigates the code and time synchronization of the cell search algorithm. In addition to code and time synchronization, frequency synchronization between the MS and the BS needs to be achieved. The cell search considered in this research is initial cell search. There is another cell search called target cell search, which needs to be performed during a call, and when an MS is in motion and moves from one cell to another. VLSI implementations to perform target cell search efficiently need to be investigated. Kiessling and Mujtaba (2002) suggest performance enhancements to the W-CDMA initial cell search algorithm. The authors consider the advantages of over sampling and passing multiple candidates in the cell search stages, instead of one candidate to reduce the cell search time.

Passing multiple candidates in each of the stages will reduce the cell search time, but increase the design complexity and hardware utilization. The improved cell search time at the expense of increased hardware utilization needs to be studied. The results presented in this chapter show that for an AWGN channel model in a high signal-to-noise ratio environment, the Improved CSD achieves faster synchronization at lower hardware complexity in comparison to the 3GPP-comma free CSD. Future work needs to investigate how the Improved CSD compares with the 3GPP-comma free CSD under multi-path channel conditions.

REFERENCES

- 3GPP RAN TS 25.211 v4.0.0. (2001-2003). *Technical specification group radio access network: Physical channels and mapping of transport channels onto physical channels (FDD)*. Retrieved from www.3GPP.org (Release 4).
- 3GPP RAN TS 25.213 v4.0.0. (2001-2003). *Technical specification group radio access network: Spreading and modulation (FDD)*. Retrieved from www.3GPP.org (Release 4).
- 3GPP RAN TS 25.214 v4.0.0 (2001-2003). *Technical specification group radio access network: Physical layer procedures (FDD)*. Retrieved from www.3GPP.org (Release 4).
- Amira, A., Bouridane, A., Milligan, P., & Roula, M. (2001). Novel FPGA implementations of Walsh-Hadamard transforms for signal processing. *IEEE Vision, Image and Signal Processing*, 148(6), 377-383.
- Dahlman, E., Beming, P., Knutsson, J., Ovesjö, F., Persson, M., & Roobol, C. (1998, November). W-CDMA: The radio interface for future mobile multimedia communications. *Proceedings of the IEEE 1998 Vehicular Technology Conference* (pp. 1105-1118).
- Higuchi, K., Sawahashi, M., & Adachi, F. (1997, May). Fast cell search algorithm in DS-CDMA mobile using long spreading codes. *Proceedings of the IEEE 1997 Vehicular Technology Conference* (pp. 1430-1434), Phoenix, AZ.
- Holma, H., & Toskala, A. (2000). *W-CDMA for UMTS: Radio access for third-generation mobile communications*. New York: John Wiley & Sons.
- Kiessling, M., & Mujtaba, S. A. (2002, May). Performance enhancements to the UMTS (W-

- CDMA) initial cell search algorithm. *Proceedings of the IEEE International Conference on Communications* (Vol. 1, pp. 590-594).
- Li, C.-F., Sheen, W.-H., Ho, J.vJ.-S., & Chu, Y.-S. (2001). ASIC design for cell search in 3GPP W-CDMA. *Proceedings of IEEE VTC 2001* (Vol. 3, pp. 1383-1387).
- Nayak, S.vS., & Meher, P. K. (1999). High throughput VLSI implementation of discrete orthogonal transforms using bit-level vector-matrix multiplier. *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, 46(5), 655-658.
- Nystrom, Jamal, K., Wang, Y.-P. E., & Esmailzadeh, R. (1998, October). Comparison of cell search methods for asynchronous wideband CDMA cellular system. *Proceedings of the IEEE International Conference on Universal Personal Communication*, Florence, Italy.
- Ojanperä, T., & Prasad, R. (1998). An overview of air interface multiple access for IMT-2000/UMTS. *IEEE Communications Magazine*, 36(September), 82-95.
- Peterson, R. L., Ziemer, R. E., & Borth, D. E. (1995). *Introduction to spread spectrum communication*. Englewood Cliffs, NJ: Prentice-Hall.
- Popovic, B.M. (1999). Efficient Golay correlator. *IEEE Electronics Letters*, 35(17), 1427-1428.
- TSGR1-2/99. (1999). *Synchronization channel with cyclic hierarchical sequences*. TSG-RAN Working Group1 Meeting 2, Nortel Networks.
- TSGR1-6/99. (1999). *New downlink scrambling code grouping scheme for UTRA/FDD*. TSG-RAN Working Group1 Meeting 6.
- TSGR1-7/99. (1999). *A modified generator for multiple-scrambling codes*. TSG-RAN Working Group1 Meeting 7, Siemens.
- TSGR1-554/99. (1999). *Generalized hierarchical Golay sequence for PSC with low complexity correlation using pruned efficient Golay correlators*. TSG-RAN Working Group1 Meeting 5, Siemens and Texas Instruments.
- Viterbi, A. J. (1995). *CDMA: Principles of spread spectrum communication*. San Francisco: Addison-Wesley.
- Wang, Y.-P. E., & Ottosson, T. (2000). Cell search in W-CDMA. *IEEE Journal of Select Areas in Communications*, 18(8), 1470-1482.
- Xilinx. (2000). *The programmable logic data book 2000*.

Chapter XIV

Turbo Equalizer: A Solution to 4G Mobiles

Priyatamkumar

B.V.B. College of Engineering and Technology, Karnataka, India

R. M. Banakar

B.V.B. College of Engineering and Technology, Karnataka, India

B. Shankaranand

National Institute of Technology Surathkal Karnataka, India

ABSTRACT

Turbo codes exhibit excellent performance gains. Turbo equalization is an iterative equalization and decoding technique. It can achieve equally impressive performance gains for communication systems. Turbo codes are send digital data over channels that require equalization, i.e. those which suffer from inter-symbol interference (ISI). Turbo equalizers have been shown to be successful in mitigating the effects of inter-symbol interference introduced by partial response modems and by dispersive channels for code rates of $R > 1/2$. The performance of iterative equalization and decoding (IED) using an M-BCJR equalizer is analyzed. Bit error rate (BER), frame error rate simulations and extrinsic information transfer (EXIT) charts are used to study and compare the performances BCJR equalizers on precoded and non-precoded channels. We predict the BER performance of Turbo equalizer using the M-BCJR equalizer from EXIT charts and explain the discrepancy between the observed and predicted performances.

INTRODUCTION

Fourth-generation (4G) mobile systems are expected to provide global roaming across different types of wireless and mobile networks. Communication may be from satellite to mobile networks and to wireless local area networks

(WLANs). The main objective of 4G is to overcome the shortcomings and limitations of third-generation (3G) systems, prime amongst which is the issue of available bandwidth. In general, 4G network architecture includes three basic areas of connectivity: PANs (e.g., Bluetooth), WANs (e.g., IEEE 802.11), and

cellular connectivity. In wireless networks, quality of service (QoS) refers to the measure of the performance for a system reflecting its transmission quality and service availability (e.g., 4G is expected to have at least a reliability of 99.99%). Supporting QoS in 4G networks will be a major challenge.

The term 4G is used broadly to include several types of broadband wireless access communication systems, including cellular telephone systems. One of the terms used to describe 4G is MAGIC—Mobile multimedia, Anytime anywhere, Global mobility support, Integrated wireless solution, and Customized personal service. The vision of 4G wireless/mobile systems are of broadband access, seamless global roaming, and Internet/data/voice communication. The 4G system provides facilities to integrate terminals, networks, and applications to satisfy the increasing user demands.

The 4G mobile networks are being developed with two main objectives. One of these objectives is to overcome the shortcomings and limitations of 3G, prime amongst which is the issue of available bandwidth. 4G systems are expected to offer a speed of over 100 Mbps in stationary mode and an average of 20 Mbps for mobile stations, reducing the downlink time of graphics and multimedia components by more than 10 times, compared to currently available 2 Mbps on 3G. The second main objective behind 4G development is to make good use of the achievements in the area of wireless technology. Currently, the 4G system is a research and development initiative based upon 3G, which is having trouble meeting its performance goals. The challenges for development of 4G systems depend upon the evolution of different underlying technologies, standards, and deployment.

Features of 4G

There are some features that are expected to be supported by 4G networks, including:

1. **High Usability and Global Coverage:** 4G networks are expected to fulfill the anytime, anywhere, and any technology requirement.
2. **Broadband Connectivity and QoS:** 4G networks provide higher bandwidths up to 100 Mbps to support multimedia services. End-to-end QoS is required.
3. **High Network Capacity:** 4G network capacity should be at least 10 times that of a 3G network.
4. **Packet-Switched Network:** 4G networks are expected to be entirely packet-switched networks.
5. **Service Personalization:** In order to overcome the saturated mobile communication market, operators will seek new 4G users in widely different locations, occupations, and economic classes. So to meet demands of these diverse users, service providers should design personal and highly customized services for them.

Limitations of 4G

Although the concept of 4G communications shows much promise, there are still limitations that must be addressed. One major limitation is operating area. Although second-generation (2G) networks are becoming more ubiquitous, there are still many areas that are not served. Rural areas and many buildings in metropolitan areas are not being served well by existing wireless networks. This limitation of today's networks will carry over into future generations of wireless systems. Another limitation is cost. The equipment required to implement a next-generation network is still very expensive. Carriers and providers have to plan carefully to make sure that expenses are kept realistic. One technique currently being implemented in Asian networks is a pay-per-use model of services. This model will be difficult to implement in the countries where the public is used to a service-for-free model (e.g., the Internet).

TURBO CODES

Over the past decade and a half, there has been an enormous amount of research effort dedicated to the analysis of iterative decoding algorithms and the construction of iteratively decodable codes or turbo codes that approach the Shannon limit. Turbo codes were introduced in 1993. The results reported sent a shockwave throughout the research community.

Error-control codes (ECCs), or channel codes, allow for reliable transmission of digital information in the presence of noise. Through this process, an information-bearing sequence of length K , called a *message*, is mapped, or encoded, into another sequence of

length $N > K$, called a *codeword*. This encoding introduces redundancy, but it also restricts the number of possible transmitted codewords, allowing for reliable communication at a lower signal-to-noise ratio (SNR) (Raphaeli & Saguy, 2000). The codeword is then modulated and transmitted through the communications channel. The received signal is a distorted version of the modulated codeword; in particular, communications channels introduce noise, normally modeled as additive white Gaussian noise (AWGN) and inter-symbol interference (ISI), the effects of which are normally modeled by a linear filter. The receiver goal can be very clearly and concisely described: the transmitted message bits should be estimated at the receiver according to a rule that minimizes the bit error rate (BER). Assuming equally likely message bits, this rule can be implemented with a maximum-likelihood (ML) detector, which estimates each message bit so as to maximize the likelihood of observing the received signal conditioned on the message bit (Raphaeli & Saguy, 2000).

ML detectors jointly and optimally perform all receiver tasks, such as synchronization,

timing recovery, channel estimation, equalization, demodulation, and decoding. Unfortunately, the computational complexity of ML receivers is prohibitive. In some cases, such as coded systems with interleavers, an ML receiver has to consider every possible transmitted message independently. For messages of 1,000 bits, this means considering 21,000 messages, much more than the current estimate for the number of atoms in the universe (Robertson, Villebrun, & Hoeher, 1995). Until the promise of quantum computers (which theoretically could analyze all possible messages simultaneously) is realized (Williams & Clearwater, 1998), or until a better strategy is discovered, exact ML detection will remain a benchmark and an object of theoretical investigation.

Traditionally, receivers employ a suboptimal divide-and-conquer approach for recovering the transmitted message from the received signal. First, timing is estimated (Raphaeli & Saguy, 2000) and the signal is sampled. Then the *equalizer* parameters are estimated (Raphaeli & Saguy, 2000; Mohel, 1993). After that, the equalizer removes the ISI introduced by the channel (Raphaeli & Saguy, 2000), so that its output can be seen as a noise-corrupted version of the transmitted codeword. Finally, the equalizer output is fed to the channel decoder which, exploiting the beneficial effects of channel encoding, estimates the transmitted message (Robertson et al., 1995).

The divide-and-conquer approach is clearly suboptimal. Consider, for instance, the problem of channel estimation. Traditionally, the channel is estimated by transmitting a known sequence, called a training sequence (Raphaeli & Saguy, 2000; Mohel, 1993), and the received samples corresponding to the training sequence are used for estimation. However, this approach, known as *trained* estimation, ignores received samples corresponding to the information bits, and thus does not use all the infor-

mation available at the receiver. To improve performance, the channel may be estimated based on all received samples.

Following are the important facets to be considered for the channel design:

- **Channel Estimation:** Channel estimates are required by the ML equalizer, and can be used to compute the coefficients of suboptimal but lower-complexity equalizers, such as the minimum mean-squared error (MMSE) linear equalizer (LE) (Raphaelli & Saguy, 2000) or the MMSE decision-feedback equalizer (DFE) (Raphaelli & Saguy, 2000). Even though the MMSE-LE and the MMSE-DFE can be estimated directly, having the channel estimates allows us to choose which equalizer is more appropriate for the channel. For instance, in channels with deep spectral nulls, DFE is known to perform better than LE.
- **Exploitation of ECC:** Most of the existing blind-channel estimation techniques operate within the divide-and-conquer framework, ignoring the presence of ECC, and normally assuming that the transmitted symbols are independent and identically distributed (iid). This approach works well at high signal-to-noise ratio (SNR). However, the last decade has seen the discovery of powerful ECC techniques such as turbo codes and low-density parity check codes that, with reasonable complexity, allow reliable transmission at an SNR only a fraction of a dB from channel capacity. When powerful codes are used and systems operate at low SNR, blind and semi-blind estimation techniques that ignore ECC are doomed to fail.
- **Low Computational Complexity:** The per-symbol computational complexity of existing ECC-aware channel estimators is exponential in the memory of the channel.

However, in applications such as xDSL and high-density magnetic recording, the channel impulse can have tens or even hundreds of coefficients. For channels with long memory, existing ECC-aware channel estimators are prohibitively complex, which motivates the study of low-complexity techniques.

Examples abound that show that it is possible to improve the performance of the divide-and-conquer approach simply by having the receiver components cooperate through an iterative exchange of information. For instance, in turbo equalizers (which assume channel knowledge), the decoder output is used by the equalizer as *a priori* information on the transmitted symbols. This produces improved equalizer outputs, which in turn produce improved decoder outputs and so on. By iterating between the equalizer and the decoder, turbo equalizers achieve a BER much smaller than that of the divide-and conquer approach with reasonable complexity. Iterative channel estimators are another important class of iterative algorithms that perform better than their non-iterative counterparts. In these algorithms an initial channel estimate is used by a symbol estimator to provide tentative estimates of the first- and/or second-order statistics of the transmitted symbol sequence. These statistics are then used by a channel estimator to improve the channel estimates. The improved channel estimates are then used by the symbol estimator to improve the estimates of the statistics and so on.

Turbo equalizers and iterative channel estimators normally rely on the forward backward algorithm by Bahl, Cocke, Jelinek, and Raviv (BCJR) (1974) for equalization. This algorithm computes the *a posteriori* probabilities (APPs) of the channel inputs given the channel output, channel estimates, and *a priori* probabilities on

the channel inputs, assuming that the channel inputs are independent. In other words, if an ECC is present, this presence is ignored. The BCJR algorithm is well suited for iterative systems, since it can use the *a priori* information at its input to improve the quality of its output, since it computes soft symbol estimates in the form of APPs. However, its per-symbol computational complexity increases exponentially with the channel memory and hence is prohibitive for channels with a long impulse response.

Graphical models for turbo codes (and low-density parity check (LDPC) codes), together with the various iterative algorithms for decoding them, have provided substantial insights into the dramatic performance improvements achievable through their use (Williams & Clearwater, 1998; Raphaeli & Saguy, 2000; Mohel, 1993; Robertson et al., 1995). The flurry of research in related topics over the last decade has produced a number of communications and signal processing algorithms that leverage turbo decoding approaches to provide similar gains in performance for a wide array of problems. In this chapter, we discuss the turbo equalization approach to coded data transmission over inter-symbol interference (ISI) channels, with an emphasis on the basic ideas and some of the practical details. The original system introduced in Glavieux, Laot, and Labat (1997) leveraged the ideas of the turbo decoding algorithm to the related problem of equalization and decoding. We seek to provide an overview of the turbo equalization approach, with an algorithmic description and intuitive explanation of each of the steps involved in

designing such a communication system.

These basic elements are contained in the most practical communication systems and are essential components of a transmitter such that turbo equalization can be used in the receiver. The role of the encoder, which is the first block in Figure 1, is to take the binary data sequence to be transmitted as input and produce an output that contains not only this data, but also additional redundant information that can be used to protect the data of interest in the event of errors during transmission. There are a wide variety of practical methods for introducing such redundancy in the form of an error control code (ECC) (also referred to as forward error correction); however, we will assume that a convolutional code is used for our purposes. The goal of forward error correction is to protect the data from the possibility of random single-bit errors or short bursts of errors that might occur in the data stream as a result of additive noise in the transmission or receiver errors. In order to ensure that such errors appear random and to avoid long error bursts, an interleaver is used to randomize the order of the code bits prior to transmission. This process is completely reversible, and is simply mirrored in the receiver.

Finally, the permuted code bits are then converted into electrical signal levels that can be modulated either at base band or onto a carrier for transmission over a pass band channel. Such modulation could take a variety of forms in such diverse applications as wired or wireless transmission, optical communications, optical data storage, magnetic recording, or even acoustic communication systems. The

Figure 1. A typical communication system

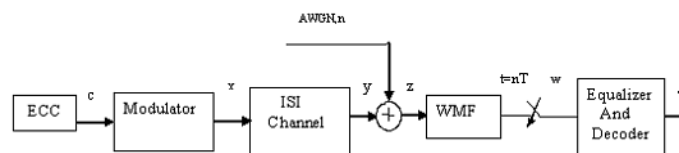
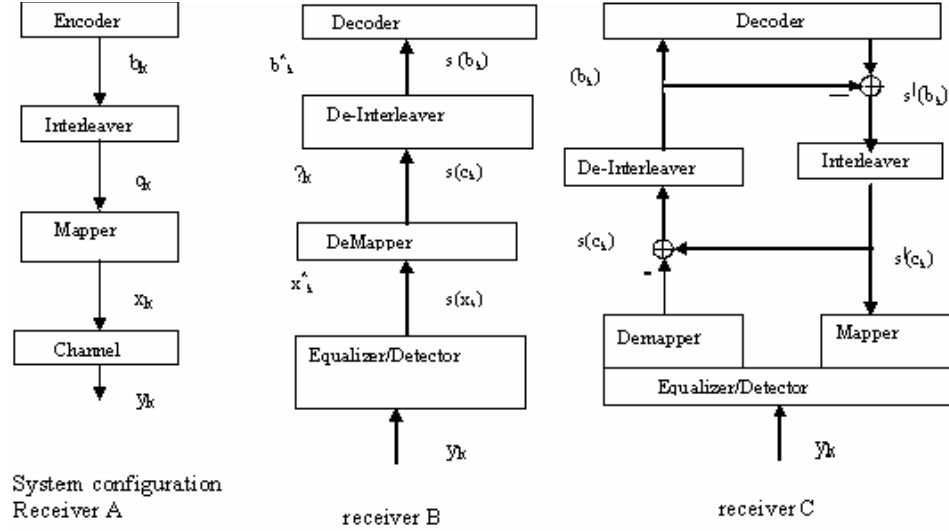


Figure 2. System configuration and three receiver structures: the optimal detector (receiver A), one-time equalization and decoding using hard or soft decisions (receiver B), and turbo equalization (receiver C)



process of mapping binary code bits into channel symbols suitable for modulation is depicted by the mapper in Figure 2.

The traditional methods of data protection used in ECC do not work when the channel over which the data is sent introduces additional distortions in the form of intersymbol interference. When the channel is limited or for other reasons is dispersive in nature, then the receiver will, in general, need to compensate for the channel effects prior to employing a standard decoding algorithm for the ECC. Such methods for channel compensation are typically referred to as channel equalization. Even when the actual transmission medium is non-dispersive, often the transmit and receive filtering that takes place in a practical system gives rise to sufficient intersymbol interference, then equalization becomes necessary. Given observations of the received data, the receiver now has essentially one task to complete: estimate the data that was transmitted. To do this optimally, in terms of minimizing the bit error rate (BER), the receiver must find the set of transmit-

ted bits that are most probable, given knowledge of the complex statistical relationship between the observations and the transmitted bits. Such a receiver, as depicted in Figure 2 as receiver A, takes into account the error control coding, the interleaving, symbol mapping, and knowledge of the channel. With so many factors involved, the resulting statistical relationship rapidly becomes difficult to manage in an efficient manner. As such, in most practical systems, receiver A is simply infeasible, as it amounts to essentially trying to fit all possible sequences of transmitted bits to the received data, a task whose complexity grows exponentially in the length of the data transmitted.

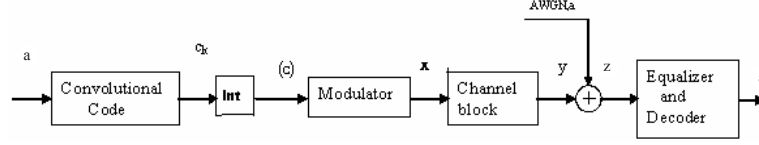
The way that most practical receivers have been designed is to first process the received observations to account for the effects of the channel and to make estimates of the transmitted channel symbols that best fit the observed data. A number of criteria for performance have been used for such equalizers, ranging from those attempting to simply invert the channel (so-called zero forcing equalizers) to linear

and nonlinear equalizers based on minimizing a mean squared error (MSE) metric to even those that are symbol-error-rate (SER) optimal by maximizing the likelihood of the observations given the channel and data model. These equalization methods constitute the first step in receiver B from Figure 2. Once the transmitted channel symbols have been estimated, they can be de-mapped into their associated code bits, de-interleaved, and then decoded using a BER optimal decoder for the ECC. The most straightforward way to implement this separate equalization and decoding process is for the equalizer to make hard decisions as to which sequence of channel symbols were transmitted, and for these hard decisions to be mapped into their constituent binary code bits. These binary code bits can then be processed with the decoder for the ECC. However, the process of making hard decisions on the channel symbols actually destroys information pertaining to how likely each of the possible channel symbols might have been. This additional “soft” information can be converted into probabilities that each of the received code bits takes on the value of zero or one, which, after de-interleaving, is precisely the form of information that can be exploited by a BER optimal decoding algorithm. Many practical systems use this form of soft input error control decoding by passing soft information between an equalizer and decoding algorithm. The remarkable performance of turbo codes made it clear that the soft information need not only flow in one direction. Once the error control decoding algorithm processes the soft information, it can, in turn, generate its own soft information, taken into account in the equalization process, creating a feedback loop between the equalizer and decoder, through which each of the constituent algorithms communicates its beliefs about the relative likelihood that each given bit takes on a particular value. This process is often termed “belief propagation” or “message passing” and has a number of impor-

tant connections to methods in artificial intelligence, statistical inference, and graphical learning theory. The feedback loop structure described here and depicted in receiver C in Figure 2 is essentially the process of turbo equalization. While the process of equalization and decoding through the feedback loop structure of receiver C is essentially complete, it is important to consider the effect that the soft information generated from one bit in one of the constituent algorithms (equalizer or decoder) will have on other bits in the other constituent algorithm. When processing soft information as an input to the equalizer or decoder, it is assumed that the soft information about each bit (or channel symbol) is an independent piece of information. This enables simple, fast algorithms to be used for each the equalizer and decoder. However, if the decoder formulates its soft information about a given bit, based on soft information provided to it from the equalizer about exactly the same bit, then the equalizer cannot consider this information to be independent of its channel observations. In effect, this would create a feedback loop in the overall process of length two—the equalizer informs the decoder about a given bit, and then the decoder simply re-informs the equalizer what it already knows. In order to avoid such short cycles in the feedback, and in hopes of avoiding local minima and limit cycle behavior in the process, when soft information is passed between constituent algorithms, such information is never formed based on the information passed into the algorithm concerning the same bit. Basically this amounts to the equalizer only telling the decoder new information about a given bit based on information it gathered from distant parts of the received signal (thanks to the interleaver). Similarly, the decoder only tells the equalizer information it gathered from distant parts of the encoded bit stream.

As a result, the iterative equalization and decoding process can continue for much itera-

Figure 3. DTTF model of transmitter



tion before cycles are introduced; this eventually limits further improvements. This process only constitutes passing “extrinsic information” between constituent decoders.

TURBO EQUALIZER

Turbo equalizer is employed in the presence of an ISI channel to counter the effects of ISI and minimize the bit error rate. It was pioneered by Douillard et al. (1995) as an application of the iterative decoding principle (Bauch, Khorrarn, & Hagenauer, 1997; Proakis & Salehi, 1994) and has since been an object of extensive research.

System Model

A typical communication system is shown in Figure 3; the data is protected by an ECC. The coded bits c_k are mapped to a symbol from the alphabet C of the signal constellation and transmitted over an ISI channel. For simplicity, we assume binary phase shift keying (BPSK) modulation—that is, $C = \{-1, +1\}$. In the receiver, the channel output z_k is passed through a whitened matched filter (WMF) and sampled every T seconds, the symbol duration. The cascade of the ISI channel with the WMF and the sampler may be equivalently treated as a DTTF. The DTTF is a finite impulse response (FIR) filter and can be represented by an L -tap delay line with coefficients f_m , $m = 0, 1, 2, L - 1$. Thus the

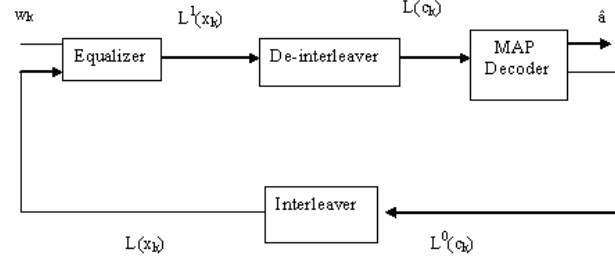
output of the DTTF at time k may be written as

$$\omega_k = \sum_{m=0}^{L-1} f_m x_{k-m} + \eta_k$$

where η_k is a sample of a white Gaussian noise process at time k . We shall refer to a communication system model in which the ISI channel, the WMF, and the sampler are replaced with the DTTF as its DTTF model.

From the analogy with SCCCs (Raphaeli & Zarei, 1997), we know that the BER can be significantly improved by iteratively passing soft information rather than hard symbol estimates between the equalizer and the decoder. However, the computational complexity of such a system increases with the number of iterations and the channel impulse response length L . To facilitate turbo equalizer, a convolutional code is used as the ECC. It is possible to obtain large interleaving gains in SCCCs if the inner code (DTTF) is recursive (Raphaeli & Zarei, 1997). To make the DTTF appear recursive, we may use a precoder as in Raphaeli and Zarei (1997). The DTTF model of the communication system is shown in Figure 4. A binary data stream, a_k , $k = 1, 2, \dots, N$ of length N is encoded by the convolutional encoder generating a multiplexed stream of coded bits, c_k , $k = 1, 2, \dots, K$ of length $K = (1/R) \cdot N$ where R is the rate of the convolutional code. The coded bits c_k are pseudo-randomly permuted by the interleaver $\Pi(\cdot)$. The interleaved stream, $\Pi(c_k)$, $k = 1, 2, \dots, K$ is BPSK modulated to obtain the sequence x_k , $k = 1, 2, \dots, K$, $x_k \in X$, which is then

Figure 4. Receiver model for turbo equalizer



convolutionally encoded by the DTTF to produce y_k , $k = 1, 2, \dots, K$, $y_k \in \xi$ where ξ is the output alphabet of the DTTF when the input alphabet is C . Since the code rate of DTTF is unity, the length of the output stream is equal to K .

Principle of Turbo Equalizer

Turbo equalizer is based on the principle of iterative decoding applied turbo equalizer to SCCCs. The receiver structure for turbo equalizer is depicted in Figure 4. Both the equalizer and the decoder are *SISO* devices. The equalizer computes the *a posteriori* probabilities (APPs), $P(x_k = x | z_1, z_2, \dots, z_K)$, $x \in X$, $k = 1, 2, \dots, K$, given K received symbols z_k , $k = 1, 2, \dots, K$, and outputs the extrinsic LLRs, $L^I(x_k)$, $k = 1, 2, \dots, K$, defined as the *a posteriori* LLR minus the *a priori* LLR. The superscript I refers to the inner *SISO* module—namely, the equalizer.

$$L^I(x_k) = \ln \left(\frac{P(x_k = +1 | z_1, z_2, \dots, z_K)}{P(x_k = -1 | z_1, z_2, \dots, z_K)} \right) - \ln \left(\frac{P(x_k = +1)}{P(x_k = -1)} \right) \quad (1)$$

The *a priori* LLR, $L(x_k)$, $k = 1, 2, K$, represents the *a priori* information about the probability that $x_k \in X$, $k = 1, 2, K$ assumes a particular value. The APP decoder provides them. In the first equalization step, the *a priori* information, $L(x_k)$, $k = 1, 2, K$, is not available and all values are assumed to be equally probable—that is, $L(x_k) = 0, \forall k$. The *a priori*

information for the decoder is obtained by de-interleaving the output extrinsic LLRs of the equalizer, $L(c_k) = \Pi^{-1}(L^I(x_k))$. Similar to the equalizer, the decoder also computes the APPs, $P(c_k = c | L(c_1), L(c_2), \dots, L(c_K))$, $c \in [0, 1]$, given the K code bit LLRs, $L(c_k)$, $k = 1, 2, \dots, K$, and outputs the extrinsic LLRs, $L^O(c_k)$, $k = 1, 2, \dots, K$, defined as the output LLR minus the *a priori* LLR. The superscript O refers to the outer decoder.

$$L^O(x_k) = \ln \left(\frac{P(c_k = +1 | L(c_1), L(c_2), \dots, L(c_K))}{P(c_k = -1 | L(c_1), L(c_2), \dots, L(c_K))} \right) - \ln \left(\frac{P(c_k = +1)}{P(c_k = -1)} \right) \quad (2)$$

The extrinsic LLRs of the decoder are interleaved to obtain the intrinsic information for the equalizer—that is, $L(x_k) = P(L^O(c_k))$, $k = 1, 2, \dots, K$. The decoder also outputs estimates of the data bits as

$$a_k = \arg \max_{b \in \{0,1\}} \{P(b_k = b | L(c_1), L(c_2), \dots, L(c_K))\} \quad (3)$$

This process is repeated several times over a block of received symbols. The BCJR algorithm and its M -variant as used in turbo equalizer are described in the next section.

SISO MODULE BASED ON THE BCJR ALGORITHM

The *SISO* module shown in Figure 5 is a four-port device that accepts at the input the se-

quences of LLRs of the information bits $\lambda^l(u)$ and the coded bits $\lambda^l(c)$, and produces at the output sequences of *extrinsic* LLRs for the information bits $\lambda^o(u)$ and the coded bits $\lambda^o(c)$.

In our description of the BCJR algorithm, we assume that the encoder trellis is time-invariant. This is a valid assumption because we are dealing with the time invariant trellises of the DTF and convolutional codes in IED. For notation, we use the trellis section of the encoder trellis shown in Figure 6. The symbol e denotes a trellis edge starting from state s_e^α and ending at state s_e^β . u_e and c_e are the respective information and the code symbols associated with the edge e .

The SISO module can operate at *bit* level or *symbol* level. Quite often, the interleaver operates at *bit* level for improved performance. This necessitates the transformation of *symbol* LLRs into *bit* LLRs and *vice versa* if the SISO module is implemented at the *symbol* level.

The Input-Output Relationships of the SISO Module

Assume that the information and code symbols are defined over a finite time index set $[1, 2, \dots, K]$. Let the operator

$$\log^*(a_j) \equiv \log \sum_{j=1}^J e^{a_j} \quad (4)$$

From the input LLRs, $\lambda^l k(u)$ and $\lambda^l k(c)$, $k = 1, 2, \dots, K$, the output *extrinsic* LLRs, $\lambda^o k(u)$ and $\lambda^o k(c)$ (c), $k = 1, 2, \dots, K$, are calculated as

$$\lambda_k^o(c) = \log^* \log \left\{ \alpha_{k-1}(s_e^\alpha) + \lambda_k^l(c_e) + \beta_k(s_e^\beta) \right\} \quad (5)$$

$$\lambda_k^o(u) = \log^* \log \left\{ \alpha_{k-1}(s_e^\alpha) + \lambda_k^l(c_e) + \beta_k(s_e^\beta) \right\} \quad (6)$$

Figure 5. Block diagram of SISO model



The LLRs calculated according to (5) and (6) are termed *extrinsic* due to the fact that the computation of $\lambda_k^o(u)$ does not depend on the corresponding input LLR $\lambda_k^l(u)$ and so it can be considered as an update. The quantities $\alpha_k(\cdot)$ and $\beta_k(\cdot)$ in (5) and (6) are obtained through *forward* and *backward* recursions, respectively, as

$$\alpha_k = \frac{\log^* \left\{ \alpha_{k-1}(s_e^\alpha) + \lambda_k^l(u_e) + \lambda_k^l(c_e) \right\}}{\log^* \left\{ \log^* \left\{ \alpha_{k-1}(s_e^\alpha) + \lambda_k^l(u_e) + \lambda_k^l(c_e) \right\} \right\}}, k = 1, 2, \dots, K-1 \quad (7)$$

SISO

assuming that the encoder starts and ends in state 0. If the encoder ends in an unknown state, the initial values for the backward recursion are chosen as $\beta_k(s) = \alpha_k(s), \dots, \forall s$

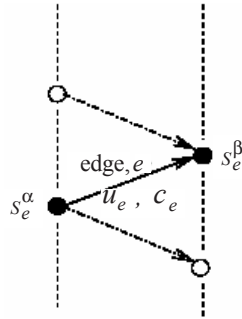
$$\beta_k(s) = \left\{ \frac{\log^* \left\{ \beta_{k+1}(s_e^\beta) + \lambda_{k+1}^l(u_e) + \lambda_{k+1}^l(c_e) \right\}}{\log^* \left\{ \alpha_k(s_e^\alpha) + \lambda_{k+1}^l(u_e) + \lambda_{k+1}^l(c_e) \right\}} \right\}, k = 1, 2, \dots, K-1 \quad (8)$$

with initial values $\alpha^o(s) = 0, s = S_0$ and $\beta_k(s) = 0, s = S_k$

- ∞ , otherwise ∞ , otherwise

The denominators in (7) and (8) are normalization terms which help avoid numerical problems arising out of finite precision. The $\log^*$ operator may be simplified

Figure 6. Encoder trellis section defining notation for description of the SISO algorithm of the input LLR based on the code constraints and the information provided by all homologous symbols in the sequence except the one corresponding to the same symbol interval



$$\log^*(a_j) = \max_j(a_j) + \delta(a_1, a_2, \dots, a_j) \quad (9)$$

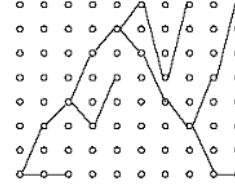
where $\delta(a_1, a_2, \dots, a_j)$ is a correction term that can be computed recursively using a single-entry lookup table (Robertson et al., 1995). This simplification significantly decreases the computational complexity of the BCJR algorithm at the expense of slight performance degradation.

Inter-Conversion between Symbol and Bit Level LLRs

Inter-conversion operations between *symbol* and *bit* level LLRs are necessitated by the presence of a bit-interleaver. These operations assume that the bits forming a symbol are independent. Suppose $\mathbf{u} = [u_1, u_2, \dots, u_m]$ is a symbol formed by m bits. The extrinsic LLR λ_j of the j the bit u_j within the symbol $\mathbf{u}$ is obtained as

$$\lambda_j^o(u) = \log_{u, u_e=1}^* [\lambda^0(u) + \lambda^1(u)] - \log_{u, u_e=0} [\lambda^0(u) + \lambda^1(u)] - \lambda_j^l(u) \quad (10)$$

Figure 7. Idealized computation pattern in the M-BCJR algorithm on an 8-state trellis



Conversely, the extrinsic LLR of the symbol $\mathbf{u}$ is obtained from the extrinsic LLRs of its component bits u_j as

$$\lambda(u) = \sum_{j=1}^m \lambda_j u_j \quad (11)$$

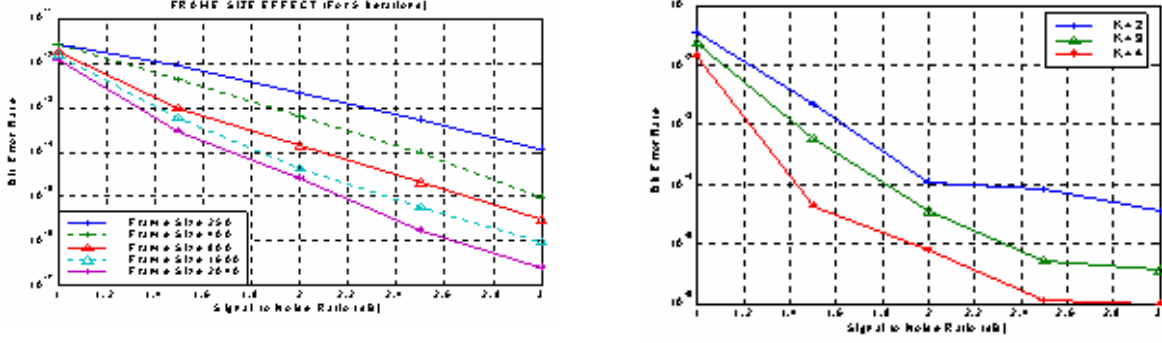
M-BCJR Algorithm

The M -BCJR algorithm (Raphaeli & Zurai, 1997) is a reduced-complexity variant of the BCJR algorithm and is based on the M -algorithm, a reduced-search trellis decoder. The reduction in complexity is achieved by retaining only the M -best paths in the forward recursion at each time instant. In the calculation of α_k through forward recursion on α_{k-1} , only the M largest components are used; the rest of them are set to an LLR of $-\infty$, and the corresponding states are thus declared dead. The backward recursion is executed only via forward recursion. In Figure 7, we show an example of M -BCJR computation pattern for $M=2$.

PERFORMANCE ANALYSIS OF M-BCJR EQUALIZER

The performance of the M -BCJR equalizer is studied and contrasted with that of the BCJR equalizer on a variety of ISI channels (precoded and non-precoded) with the help of BER and FER simulations and EXIT charts (Tenbrink,

Figure 8. Frame size effect and effects of puncturing on BER



2001). The ISI channels are modeled as convolutional codes (DTTFs). In all our simulations, the channel is assumed to be static, and its coefficients f_m , $m = 0, 1 \dots L - 1$ where L is the length of the channel impulse response are perfectly known. Each of the channel coefficients has a power and is $p_m = |f_m|^2$, $m = 0, 1, \dots, L - 1$ normalized such that the total power

$$\sum_{m=0}^{L-1} p_m = 1$$

In this chapter, we represent an ISI channel by its coefficients $[f_0, f_1, \dots, f_{L-1}]$. We also investigate the performance of the M -BCJR equalizer on precoded channels because precoding improves the asymptotic performance of SCCs (Benedetto, Montorsi, Divsalar, & Pollara, 1998). Precoding is achieved by appropriately processing the interleaved bits stream prior to passing it through the DTTF.

Performance of the M -BCJR Equalizer

The larger the frame size, the bigger the S-window. Therefore, it will produce larger distance by using an interleaver. The correlation between the two adjacent bits will become smaller. Hence the decoder gives better performance. The simulation results verified this conclusion. However, since turbo code is a

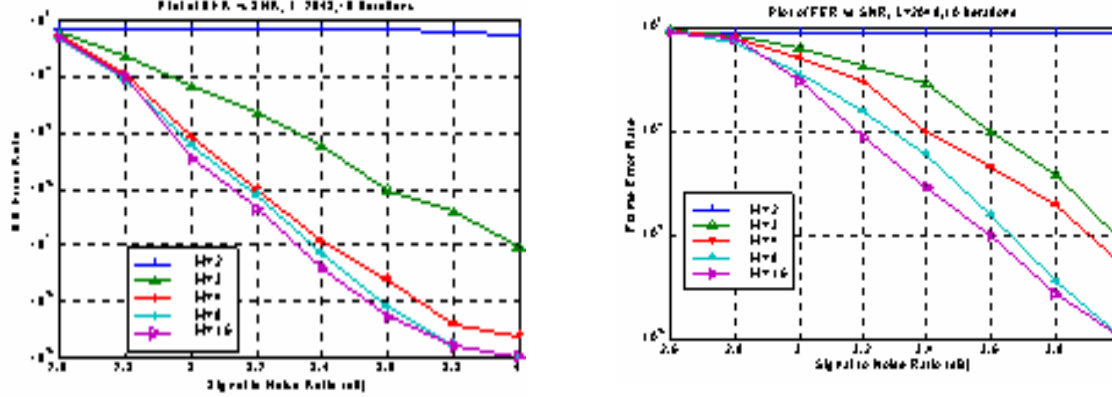
block code, it causes a time delay before getting the complete decoding output. Increasing the frame size also increases the delay time. Figure 8 shows the BERs of turbo code under static channel with the code rate=1/3, iteration=3, and frame size $L = 1024$ bits.

When the code rate is decreased, more bits have to be punctured. The bandwidth requirement is also decreased. However, some information is lost. This means that the performance of the turbo code will also degrade in general. Figure 8 shows the effects of the punctuation on BER. The higher the code rate, the lower the BER. In the simulation, decode iteration=5, frame size=1024, and uncorrelated AWGN environment is applied. The three curves correspond to code rate=1/2, 1/3, and 1/4 respectively.

In Figure 9, we show the BER and FER performance of IED for the unprecoded channel $[\sqrt{0.45}, \sqrt{0.25}, \sqrt{0.15}, \sqrt{0.10}, \sqrt{0.05}]$ with $[1, 23/35]_8$ as the outer code using 10 iterations and $N=2048$. The loss in performance of the M -BCJR equalizer as compared to the BCJR equalizer is unexpectedly small for non-precoded ISI channels.

The BER and FER curves of the $M=8$ BCJR algorithm are almost overlapping with those of the full BCJR which operates on the whole 16-state trellis. When $M=4$, the loss in performance is only 0.05 dB at a BER of 10^{-5} . For

Figure 9. BER & FER: non-precoded $[\sqrt{0.45}, \sqrt{0.25}, \sqrt{0.15}, \sqrt{0.10}, \sqrt{0.05}]$, $[1, 23/35]_8$



$M=3$, the loss is 0.25 dB at a BER of 10^{-4} . For $M=2$, the IED algorithm fails to evolve and does not provide any improvement in performance with iterations. As can be seen from these results for the above channel, we may use the $M=4$ BCJR equalizer with virtually no performance degradation or the $M=3$ BCJR equalizer with a very small loss in performance. This is an interesting result and suggests that the complexity of the BCJR equalizer can be reduced considerably without sacrificing its performance.

In Figure 10, we present the BER and FER simulation results over six iterations for the unprecoded $[0.5, 0.5, 0.5, 0.5]$ channel using the full BCJR and $M=3$ BCJR equalizers. An information block length of $N=2048$ was used. The full BCJR equalizer operates on the whole trellis consisting of eight states at each time instant. From the plots, we observe that the performance of the $M=3$ BCJR equalizer is almost indistinguishable from that of the full BCJR equalizer in the region of $\text{BER}=10^{-5}$ at reasonably high E_b/N_o . The performance of the $M=3$ BCJR equalizer is relatively worse at low E_b/N_o . The performance of the turbo equalizer saturates at a BER of 10^{-5} and does not improve significantly even at very high E_b/N_o . Such an early error floor is typical of SCCs in which the inner code is non-recursive (Bauch & Franz,

1998). It can be avoided by precoding the channel. The BER and FER simulation results for the precoded $[0.5, 0.5, 0.5, 0.5]$ channel using the full BCJR, $M=5$, and $M=3$ BCJR equalizers are plotted in Figure 11. In comparison with the non-precoded channel, the M -BCJR equalizer suffers significant losses in the precoded case. The performance of the $M=5$ BCJR equalizer is approximately 0.25 dB worse than the BCJR algorithm at a BER of 10^{-4} . However, this difference diminishes as we progress toward smaller BERs. For $M=3$ BCJR, the IED hardly yields any improvement in the BER performance with an increasing number of iterations. This is in stark contrast with its performance on the non-precoded channel. It is also interesting to note that although the asymptotic performance on precoded channels is better at high E_b/N_o , the non-precoded channels offer better performance during the first

little iteration. This behavior is a result of the fact that the initial reliability of a precoded channel is smaller than that of a non-precoded channel. However, as the iterations progress, the precoded channel outperforms the non-precoded channel. Also, there are no signs of an error floor at a BER of 10^{-5} , and thus the performance may improve significantly as E_b/N_o increases.

Figure 10. Simulation results: non-precoded $[0.5, 0.5, 0.5, 0.5]$, $[1, 23/35]_8$

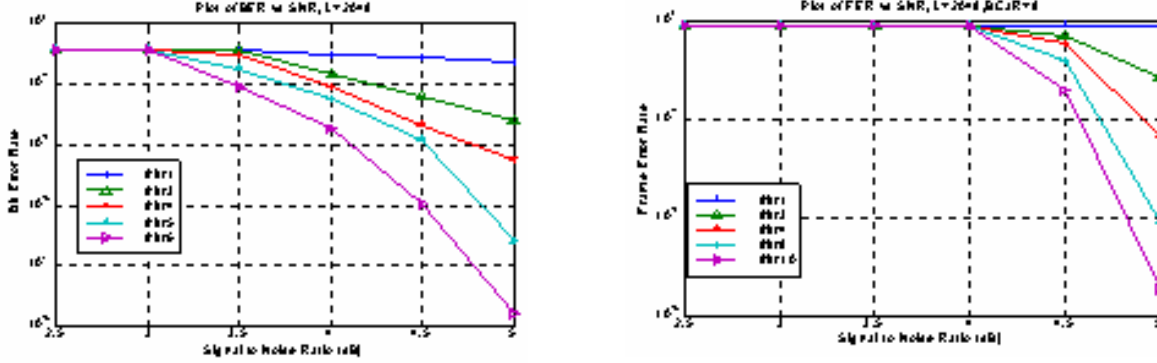
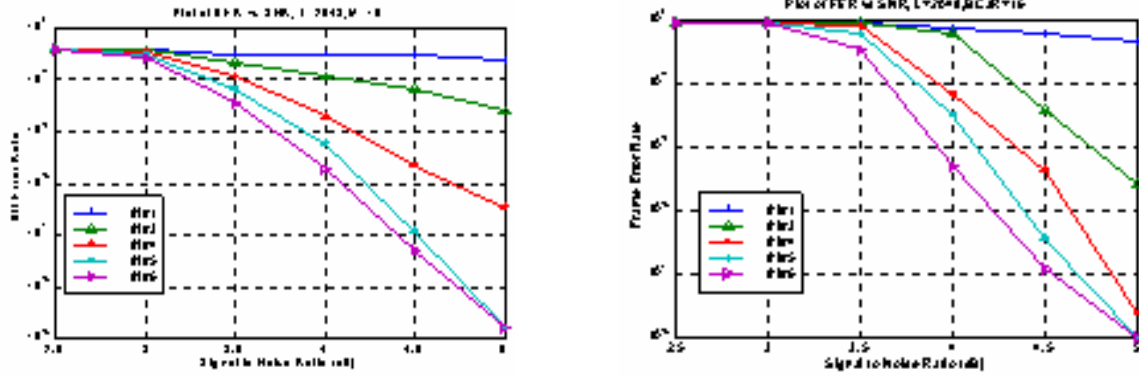


Figure 11. Simulation results: 10D precoded $[0.5, 0.5, 0.5, 0.5]$, $[1, 23/35]_8$



CONCLUSION

In this chapter, turbo code was studied in detail. The simulation results show that turbo code is a powerful error correcting coding technique in low SNR environments. It has achieved near-Shannon capacity. However, there are many factors to be considered in the turbo code design. First, a trade-off between the BER and the number of iterations must be made, for example, more iteration will get lower BER, but the decoding delay is also longer. Second, the effect of the frame size on the BER must also be considered. Although the turbo code with larger frame size has better performance, the output delay is longer. Third, the code rate is another factor to be considered. The higher coding rate needs more bandwidth. From the simulation results, it is observed that the draw-

back of the turbo code is its complexity and also the decoding time.

The focus was on understanding the behavior of the M -BCJR equalizer in IED. The performance of the M -BCJR equalizer on precoded and non-precoded channels with the help of BER, FER simulations, and EXIT charts was studied. The performance degradation of the M -BCJR equalizer as compared with the BCJR equalizer was found to be rather small. A comparison was made between the losses in achievable channel capacity of IED. Performance improved when the BCJR equalizer is replaced with the M -BCJR equalizer and an MMSE LE. It was observed that the capacity loss for the M -BCJR equalizer was very small. On the other hand, the MMSE LE suffered a substantial loss in channel capacity. It was also observed that precoding only affects the IED

performance. It does not affect the channel capacity.

Simulation results showed that the M -BCJR algorithm suffers significant losses in the case of simple convolutional decoders. This contrasting behavior of the M -BCJR algorithm in the cases of ISI channels and convolutional codes has been explained. It can be attributed to the metrics computed during the forward and backward recursions. The M -BCJR equalizer has much larger variance than in the case of a convolutional code. The larger variance of the metrics in the M -BCJR equalizer makes the algorithm less sensitive to the paths discarded.

REFERENCES

- Arnold, D., & Loeliger, A. (2001, June). On the information rate of binary input channels with memory. *Proceedings of the IEEE International Conference on Communications* (Vol. 9, pp. 2692-2695).
- Bahl, L., Cocke, J., Jelinek, F., & Raviv, J. (1974, March). Optimal decoding of linear codes for minimizing symbol error rate. *IEEE Transactions on Information Theory*, 20, 284-287.
- Banakar, R. M. (2004, July). *Low power design methodology for turbo encoder and decoder*. PhD Thesis, IIT Delhi, India.
- Bauch, G., & Franz, V. (1998, May 18-21). Iterative equalization and decoding for the GSM-system. *Proceedings of the IEEE 48th Vehicular Technology Conference*, Ottawa, Canada.
- Bauch, G., Khorram, H., & Hagenauer, J. (1997, September 30-October 2). Iterative equalization and decoding in mobile communications systems. *Proceedings of the European Personal Mobile Communications Conference*, Bonn, Germany.
- Benedetto, S., Montorsi, G., Divsalar, D., & Pollara, F. (1998, May). Serial concatenation of interleaved codes: Performance analysis, design and iterative decoding. *IEEE Transactions on Information Theory*, 44, 909-926.
- Berrou, C., & Glavieux, A. (1996, October). Near optimum error-correcting coding and decoding: Turbo codes. *IEEE Transactions on Communications*, 44, 1261-1271.
- Berrou, C., Glavieux, A., & Thitimajshima, P. (1993, May 23-26). Near Shannon limit error-correcting coding and decoding: Turbo codes. *Proceedings of the International Conference on Communications*, Geneva, Switzerland (pp. 1064-1070).
- Cover, T., & Thomas, J. (1991). *Elements of information theory*. New York: John Wiley & Sons.
- Douillard, A. Picart, M. Jézéquel, P., Didier, Berrou, C., & Glavieux, A. (1995, September-October) Iterative correction of intersymbol interference: Turbo-equalization. *European Transactions on Communications*, 6, 507-511.
- Forney, G. (1972, May). Maximum-likelihood estimation of digital sequences in the presence of intersymbol interference. *IEEE Transactions on Information Theory*, 18, 363-378.
- Gertsman, M. J., & Lodge, J. L. (1997, July). Symbol-by-symbol MAP demodulation of CPM and PSK signals on Rayleigh flat-fading channels. *IEEE Transactions on Communications*, 45, 788-799.
- Glavieux, A., Laot, C., & Labat, J. (1997, September). Turbo equalization over a frequency selective channel. *Proceedings of the International Symposium on Turbo Codes*, Brest, France (pp. 96-102).
- Hagenauer, J., Offer, E., & Papke, L. (1994, March). Iterative decoding of binary block and

- convolutional codes. *IEEE Transactions on Information Theory*, 42, 429-445.
- Heegard, & Wicker, S. (1999). *Turbo coding*. Boston: Kluwer Academic.
- Hirt, W., & Massey, J. (1988, May). Capacity of the discrete-time. *Gaussian channel with intersymbol interference*, 380-388.
- Kavcic. (2003, November). On the capacity of Markov sources over noisy channels. *Proceedings of the Global Communications Conference* (Globecom).
- Lee, & Messerschmitt, D. G. (1994). *Digital communication* (2nd ed.). Boston: Kluwer Academic.
- Li, Y., Vucetic, B., & Sato, Y. (1995, May). Optimum soft-output detection for channels with inter symbol interference. *IEEE Transactions on Information Theory*, 41, 704-713.
- Lin, S., & Costello, J. J. (1983). *Error control coding*. Englewood Cliffs, NJ: Prentice-Hall.
- Moher, M. (1993, November 29-December 2). Decoding via cross-entropy minimization. *Proceedings of the IEEE Global Telecommunications Conference*, Houston, TX (pp. 809-813).
- Poor, H. V. (1994). *An introduction to signal detection and estimation* (2nd ed.). Berlin: Springer-Verlag.
- Proakis, J., & Salehi, M. (1994). *Communication systems engineering*. Upper Saddle River, NJ: Prentice-Hall.
- Rabiner, L. (1989, February). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of IEEE* (Vol. 77, pp. 257-286).
- Raphaeli, D., & Saguy, A. (2000, September). Linear equalizers for turbo equalization: A new optimization criterion for determining the equalizer taps. *Proceedings of the 2nd International Symposium on Turbo Codes* (pp. 371-374), Brest, France.
- Raphaeli, D., & Zarai, Y. (1997, November 3-8). Combined turbo equalization and turbo decoding. *Proceedings of the Global Telecommunications Conference*, Phoenix, AZ (pp. 639-641).
- Robertson, P., Villebrun, E., & Hoeher, P. (1995, June). A comparison of optimal and suboptimal MAP decoding algorithms operating in the log domain. *Proceedings of the IEEE International Conference on Communications*, Seattle, WA (pp. 1009-1013).
- Souvignier, T., Friedman, A., Öberg, M., Siegel, R. E. S. P. H., & Wolf, J. (1999, June 6-10). Turbo decoding for pr4: Parallel versus serial concatenation. *Proceedings of the International Conference on Communications*, Vancouver, Canada.
- Ten Brink, S. (2001, October). Convergence behavior of iteratively decoded parallel concatenated codes. *IEEE Transactions on Communications*, 49, 1727-1737.
- Tüchler, M., Singer, A., & Kötter, R. (2002, March). Minimum mean squared error (MMSE) equalization using priors. *IEEE Transactions on Signal Processing*, 50, 673-683.
- Weiss, Bettstetter, C., & Riedel, S. (2001). Code construction and decoding of parallel concatenated tail-biting codes. *IEEE Transactions on Information Theory*.
- Wicker, S. B. (1995). *Error control systems for digital communication and storage*. Englewood Cliffs, NJ: Prentice-Hall.
- Williams, P., & Clearwater, S. H. (1998). *Explorations in quantum computing*. New York: Springer-Verlag.

Chapter XV

Administration of Wireless Software Components

Franck Barbier

PauWare Research Group, France

Fabien Romeo

PauWare Research Group, France

ABSTRACT

Software components embedded in mobile and wireless devices, as ordinary components deployed in distributed systems, have to be managed in order to recover faults, to trace and analyze behaviors, to enable business services such as online maintenance, customer practice understanding and so on. Despite the existence of management standards and platforms, the implementation of management facilities inside components as well as the possibility to access and operate these facilities by means of appropriate interfaces (a configuration interface for instance in order to instrument dynamical re-configuration) are not actually available. In this scope, this chapter discusses and provides a design method and an associated Java library in order to have manageable and self-manageable components specific to mobile and wireless environments.

INTRODUCTION

A major trend concerning the mobile system industry is the need for designing software applications as assemblies of reusable components, such as Java components in J2ME environments and C# components in Windows CE environments. Components are interconnected

through well-defined interfaces, while their implementation remains hidden. This results in increased reuse, easy isolation of faults, and overall improvement of quality and reliability. Furthermore, it also enables components to be independently deployed by third parties. Deployment occurs on various devices such as mobile phones, personal digital assistants

(PDAs), set-top boxes, smart cards, and so on. Owing to the fact that mobile deployment environments are different from development environments, abnormal behaviors and/or misuses occur. Consequently, some deferred assessment is only possible if components have been endowed with remote administration capabilities, including state and behavior supervision, even control, at the time components face unstable execution contexts.

From a business perspective, such a global approach is a way for supporting and thus for offering better customer services: better start-up processes, analysis of deficiencies, possible corrections through (re)configurations, marketing-based studies of common practices and expectations, and so forth. All of the well-known qualities of the component-based development technology (National Coordination Office for Information Technology Research and Development, 2001; Szyperski, Gruntz, & Murer, 2002) seem promising to achieve the necessary flexibility and adaptability imposed by the mobile systems' market.

The purpose of this chapter is to propose appropriate concepts, techniques, and tools to manage component behaviors and their associated interactions within wireless devices. Management activities stress component behavior tracing and possible dynamic (re)configuration in order to ensure actual supervision and control of mobile devices. Some of the contribution comes from the results of the Component+ European project (www.component-plus.org) in which the idea of Built-In Test (Barbier, 2005) has been formalized, developed, and put into practice for multimedia software components.

This chapter extends some of the ideas relating to the BIT technology that initially does not support administration facilities. We especially deal with the notion of "wireless software component." Even if this expression may *a*

priori not make sense, we here mean software components deployed in wireless systems. The chapter is organized in three main parts. The first part reviews the current relation between CBSE and software for mobile systems. After justifying an interesting potential synergy between the two domains, a discussion about component management architecture standards as Java Management eXtensions or JMX (Sun Microsystems, 2002; Kreger, Harold, & Williamson, 2003) is proposed. This first part ends with a focus on UML 2 in particular and executable modeling language in general, in order to answer the question: What could be a suitable technical framework for wireless component administration?

While the second part of the chapter exposes our contribution in terms of executability of the UML 2 notation and its inherent implementation for component administration, the third and final part deals with a fully implemented case study—a home automation system, including a complex wireless software component corresponding to a programmable thermostat. The state machine diagram of this component is in the Appendix. For illustration purposes, screen shots are offered, especially those relating to the programmable thermostat's management activities in Web browsers.

SOFTWARE FOR MOBILE AND WIRELESS DEVICES: AN OVERVIEW

Nowadays, a great diversity of mobile devices is offered to consumers in different market segments such as telephony, digital interactive television, home automation, and automotive industry. Each device category uses specific hardware architectures and equipment to better fit customer requirements. For instance, PDAs use more powerful processors than mobile

phones and also provide larger screens; and car navigation systems, for instance, use voice recognition, while cheaper devices often support simpler software functionality. Despite the existence of varied functionality, software for wireless devices primarily benefits from being constructed on the top of recognized and stable platforms such as J2ME in the Java world. Next, the existence of norms permits laying down assumptions for programmers so that their code does not deal with a large number of technical development/deployment contexts. For user interfaces of PDAs and cell phones, for instance, the MIDP (mobile information device profile) norm (Bloch & Wagner, 2003) is in this scope a relevant framework.

However, new software paradigms such as autonomous computing may make software components not aware of their computing environment (e.g., new possible interactions with unknown components, expected dynamic adaptations). Since this problem cannot be tackled at development time, but only when the system is deployed, there is a need for extra code that may anticipate such potential challenges. In this line of reasoning, we have defined the BIT technology that refers to testing code that remains at runtime. As for administration, it refers to the possibility for ruling such embedded code remotely, because of the need for administration policies involving not only a single component but a system of components. In the case of mobile devices, overheads caused by this extra code raise performance problems and thus create the necessity of an adequate management architecture presented in this chapter.

WIRELESS SOFTWARE COMPONENTS

The component technology (Szyperski et al., 2002) has been proven mature for numerous

application domains such as Web-based and business applications, and real-time systems. However, its usage has been evolving slowly in the field of mobile systems.

Components indeed run in large infrastructures involving different server types, more or less normalized protocols, and sophisticated services (e.g., timer services, transaction management services), all of these being often incompatible with hardware features of mobile systems that offer a reduced set of possibilities for software deployment. Software components may thus exist within wireless devices, but many interesting system-based services have to be externalized. One may for instance imagine transaction management facilities that can only be implemented on a server-side base with appropriate communication. Applying roll-back actions for a wireless software component then becomes not so easily compared to an Enterprise JavaBeans in an ordinary J2EE server. This especially results from the need for transaction management-oriented code that may clutter, even damage, any user-oriented functioning in a wireless device. However, in the field of distributed applications, the need for management of not only component units, but component assemblies, makes software in mobile systems a piece in a puzzle but not a standalone running machine—a wireless device is no more than a node in a distributed application. At least, virtual links have thus to be created with components that run on other nodes (called remote components later in this chapter), but have close collaboration with those on mobile systems.

Wireless software components are in essence reactive systems in the sense that their behavior mostly consists of processing a lot of events of different types (incoming communication flow, user interface interaction, etc.). The high frequency of events and the large spectrum of interpretation contexts show that

state machines are good candidates for behavior specification. More generally, building software for mobile systems with design languages that have interesting features such as (i) being standards, (ii) supporting executability (Mellor & Balcer, 2002), and (iii) supporting compositionality (Bock, 2004), permits one to weave—more clearly and more rigorously—wireless components with server-side components with which they have to collaborate.

ADMINISTRATION

The idea of administration (the term management has been used interchangeably) stems from the network domain in which network element attributes and behaviors have to be supervised and, in case of failure, they also have to be driven so that communication occurs as well as possible: availability of damaged modes, correction actions as switching communication flows, and so forth.

The need for standardization has generated dedicated administration supports, namely dedicated protocols including SNMP (Simple Network Management Protocol) or CMIP (Common Management Information Protocol) and, in the world of Java and software components, the Java Management eXtensions framework.

Administration covers two main activities: supervision (e.g., *in-situ* testing) and monitoring (e.g., dynamical (re)configuration). In the Java global context, network elements that are “resources” that are viewed within application top layers as software components are often named managed objects.

Java Management eXtensions

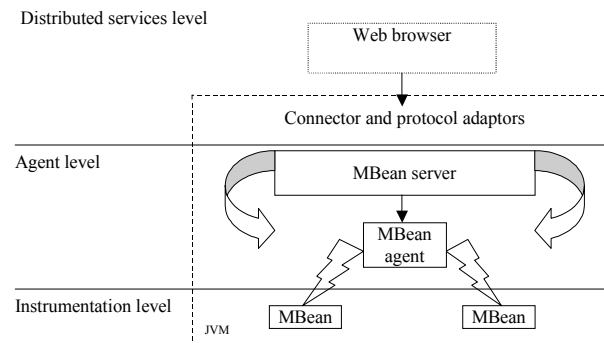
JMX has been specified and implemented to have an open standard for managing resources based on an abstract approach. JMX high level

of abstraction does not rely on specific protocols or other platform-dependent properties. Moreover, the wireless nature of resources and software components may be hidden without changing administration strategies and policies. We here see a concrete advantage of incorporating software components into wireless devices: an uniformed access to resources that is independent of running contexts, as for instance Bluetooth-based protocols and, more generally, any coercive communication mechanism imposed by wireless systems.

JMX architecture is grounded on three main categories of software components (Figure 1):

- MBeans (Managed Beans) are the components that can be managed by JMX. They define standard interfaces in which a high or low number of functional operations (normal behavior) as well as in-house management operations (e.g., state observation and tracing) will be accessible at an administration level.
- MBean agents that are directly deployed in JMX servers are responsible for controlling their locally registered MBeans. An MBeans agent generally copes with a set of MBeans in relation to predefined JMX services to ease management of MBeans. These services are common services such as timer services, but also more relevant ones such as relation services that are peculiar to compositionality management of component systems. We may also notice that JMX owns a metalevel, in the sense that MBean agents are themselves manageable entities at the Distributed services level (see Figure 1).
- Connectors are plugged into MBean agents to make them remotely available to management applications. The communication protocol is specific to a connector type that in essence aims at linking JMX to

Figure 1. JMX stereotyped architecture (Sun Microsystems, 2002)



third-party applications. Figure 1 shows the case of an HTML adaptor that connected some given agents with a Web browser.

Administration of Wireless Environments

The management of pervasive computing environments is forced by the lack of processing power of smart devices such as cellular phones. In other words, such devices rely on layered distributed systems that aim to coordinate and effectively ensure most of the expected functionality. So, even though the idea of administering wireless environments is not new, there is no standard and rigorous method for designing wireless software components so that they are really manageable.

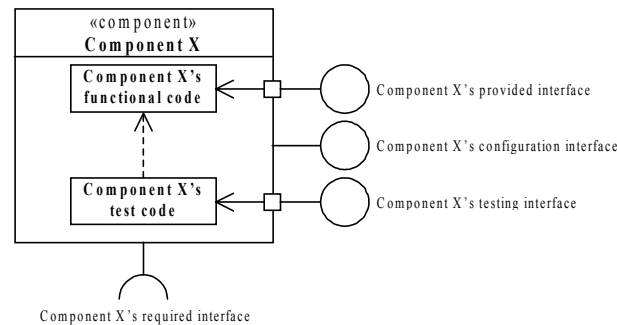
An illustration of current problems is, for instance, dynamic (re)configuration. One may imagine such a management-based use case for a mobile device in the context of a business-oriented service, which is here online maintenance. For devices equipped with the very latest technologies, including hardware and software, statistics show that return rates for reparation are close to 20%! This is exclusively due to market pressure that prompts to sell

products that are not tested enough and may even be non-mature.

Technically, the delivering of patches resulting from a mandatory, prior, and remote in-depth testing, itself based on resuming and/or on simulating special running contexts, cannot be yet easily instrumented. What indeed does “resetting an embedded component to have an initial (safe) context” or “setting up a component to an accurate context (a well-defined state)” mean?

The novelty of our approach is to design components, their internal parts especially, so that the desired administration operations may be logically described. More precisely, this includes the possibility to formally express what is a stable running context, namely forcing a given wireless component to immediately be in a set of parallel states. This also amounts to reaching a context in safely terminating any data processing, namely monitoring entry and exit operations for states. Our solution is grounded on state machine modeling languages, as well as formal execution (simulation) rules of state machines that are plugged into components. In addition, the *PauWare.Statecharts* Java library is offered to achieve all of these objectives.

Figure 2. BIT style for a wireless software component



Most of the contribution presented here is an extension of the BIT technology that initially does not address management issues. Briefly speaking, this technology advocates persistent test code in components (see Figure 2) in order to support deployment-based testing. State machines are used for instrumenting model-based testing, and above all, for having the possibility of rapidly prototyping components for assessment purposes. Another major point of BIT is to be a gate for accessing a more or less important quantity of a component's encapsulated part. For management purposes, BIT-based wireless components have required interfaces, functional (a.k.a., provided) interfaces, testing interfaces, and also configuration interfaces that are specific to management (see Figure 2). The two last sorts of interfaces are made up of services whose implementation explores and, in the spirit of administration, possibly changes a component's inside.

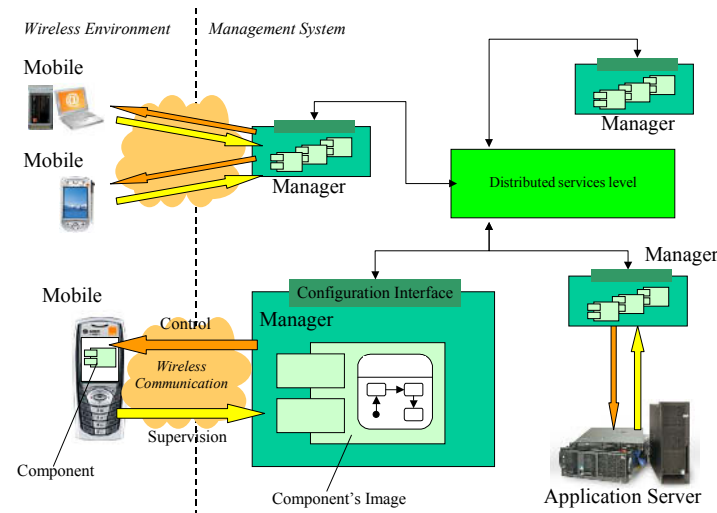
UML 2-BASED DESIGN OF WIRELESS SOFTWARE COMPONENTS

An outstanding feature of wireless components is that they have to react to many distinct types of events whose interpretation varies accord-

ing to several discrete contexts. Regarding graphic user interfaces of smart devices for instance, Horrocks (1999) advocates the use of statecharts (Harel, 1987) for designing and implementing such interfaces. On a cellular phone for example, each physical button is associated with a typed event and a precise context that leads to a specific processing at the time a given function is running under the auspices of the phone's owner.

More generally, in the scope of CBSE, the recent release (version 2) of UML (OMG, 2003a, 2003b) offers all of the necessary and sufficient integrated modeling constructs, including state machine diagrams (i.e., Harel's statecharts), sequence diagrams (useful for communication modeling), and component diagrams for the design of manageable wireless components. All of this happens in the context of the MDA/MDE initiative in which executability promotes lightweight model checking (Mellor & Balcer, 2002) and seamless implementation through model transformation: from platform-independent models as the statechart of a programmable thermostat wireless component in the Appendix to platform-dependent models that, for instance, may refer to specific properties as coercive communication as mentioned above.

Figure 3. Management architecture



A main point of UML 2 state machine diagrams is that they favor any reasoning with *abstract* states that aim at capturing the key situations and the critical phases in which mobile devices require supervision and control. Once again, in the absence of an underlying formalism, management activities cannot rely on precise representations as those offered by state machine diagrams.

TECHNICAL FRAMEWORK

Since supervision and control are dual activities involved in management, one may wonder how

both activities have to use component states. In practice, while control implies that managers change—under well-defined conditions and circumstances (see *reset* Java method below)—states of managed components, supervision consists for managers of acquiring information on current states of managed components and clusters of components. Considering that these activities have to be realized by means of wireless communication and in relation with highly constrained devices, our intention is to minimize overheads generated and sustained by our management system on the mobile side. From a business point of view, the quality of service must not be attenuated by administra-

Figure 4. Canonical behavior of a wireless component

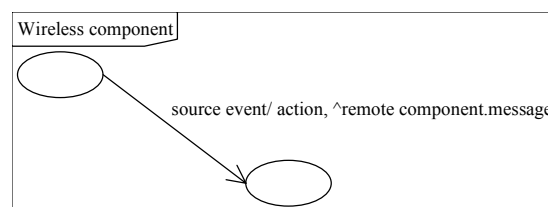


Figure 5. Common communication scraps in wireless environments

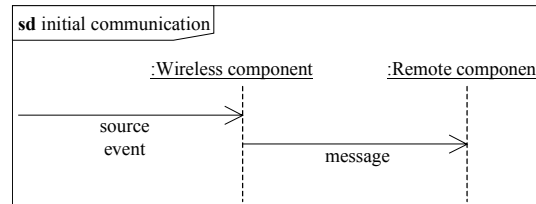
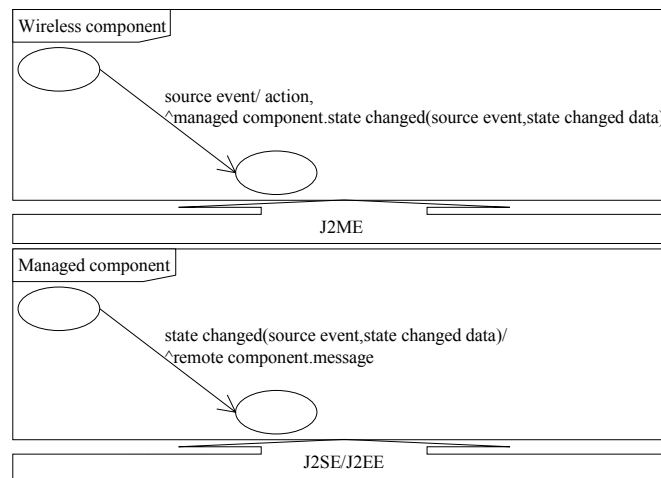


Figure 6. Overview of the technical framework



tion course. We purposefully create replications (caches or images) of managed components in the management system (see Figure 3).

Wireless Component State Machines

Figure 4 depicts the canonical behavior of an ordinary wireless software component. In other words, *source event* corresponds to a typed reaction to an external phenomenon as, for instance, a pressed device button. Numerous actions may be (in order: use of “,” for formalizing chronology in Figure 4) launched in re-

sponse to the processing of *source event*. Most of the actions are the displaying and/or the refreshing of the mobile device’s user interface.

Only two actions are drawn in Figure 4: *action* is an internal action (typically, displaying something) in the sense that it does not require any other resources than those available within the modeled device. In contrast to *^remote component.message*, *action* also does not yield any communication. In UML, an instance of the *SendSignalAction* metatype is represented by an expression similar to *^remote component.message*. That is by definition a communication unit towards a remote compo-

nent in order to create a global collaboration between several components. Such a collaboration corresponds to a strict requirement (“use case” in UML) of the designed wireless environment. This may be summarized by means of a UML Sequence Diagram (see Figure 5).

Mirroring

At this time, components deployed in mobile devices have poor autonomy. Therefore, most collaborations, like the one shown in Figure 5, are recurrent and systematic. We here mean that peer-to-peer communication, nowadays, may not be considered as a technological lock for mobile devices, but transforming such equipment into Internet “points” (i.e., Internet self-contained and autonomous nodes) is slowed down by many factors, such as hardware and platform limits, security constraints, and so forth. For instance, one cannot today run a JMX server on a mobile device for administration purposes.

Starting from this general observation, our technical framework consists of having an intermediate or mirror component, named managed component, between wireless and remote components. Moreover, managed components are exact clones of wireless components (see Figure 3). The difference between wireless components and managed components is, as indicated in their generic name, that these last ones are plugged into management environments. In Figure 6, states machines of each

type are specified, as well as a new way for communicating: We review the models in Figures 4 and 5 so that managed components may act as proxies between wireless and remote components.

In addition, since our technical framework is based on Java, we express the need for a J2SE/J2EE architecture in order to have at one’s disposal administration facilities. On the other hand, J2ME serves as the software platform for wireless components. The idea is simple: administrating wireless components occurs through the administration of perfect clones.

Communication is thus revised in Figure 6 so that exchanges with remote components are replaced in wireless components by notifications of state changes carrying all of the necessary data which allows managed components to rule their automaton in symbiosis with that of their source. Clearly, the automaton of a wireless component is implemented twice: within the mobile device in which it acts as a driver, and through the automaton of the managed component. It always possesses the same shape in the sense that managed components imitate their sources. Furthermore, Figure 6 shows that managed components serve as delegates between wireless and remote components: *^remote component.message* now appears in managed components’ automata instead of in those of wireless components.

While the automaton on the top of Figure 6 supplies the emission of the *state changed* message, the other one, beneath, shows the

Figure 7. Common communication scraps in administrated wireless environments

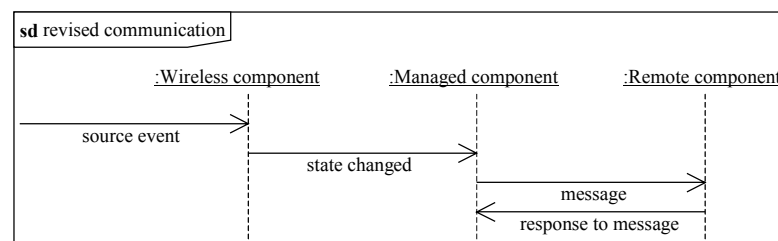
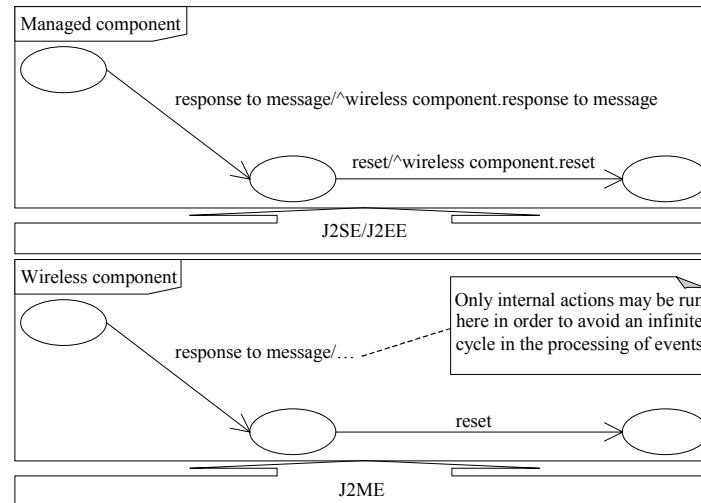


Figure 8. Inverse direction of communication for control purposes



consequential reception. As a result, both automata are “equivalent,” meaning that they are not physically synchronized, but they go through the same lifecycle. The unique problem relating to such an interval is that control activities have to be done with care. However, the original communication mechanism of statecharts is broadcast communication (Harel, 1987) that by definition makes no assumption about receivers’ status. That is the case here when possible responses go back to wireless components: any state is acceptable. The broadcast communication mechanism is supported by the *PauWare.Statecharts* library.

Control

Managed components, as members of distributed applications, are subject to requests coming from varied components. In Figure 7, the scenario shows that an event called *response to message* must be processed within the managed component’s statechart that has previously sent *message*. From a functional point of

view, *response to message* is destined to the managed component’s source: the wireless component is a replication.

Besides, managed components in essence receive administrative requests (e.g., *reset* in Figure 8) that also have to be propagated to wireless components. As a result, models in Figure 8 sum up both kinds of reaction.

Design Method

The generic micro-architecture specific to our notions of wireless component and managed image may be synthesized by means of a UML Component Diagram (see Figure 9). Wireless components provide two kinds of interfaces. The first one is for local events (*Wireless component local provided interface*) as, for instance, keyboard inputs (*source event* in Figures 4, 5, 6, and 7). The second one is for requests coming from outside (*Wireless component remote provided interface*) that must together be implemented by wireless and managed components (example: *response to mes-*

Figure 9. Assembly pattern between wireless and managed components (UML 2 Component Diagram)

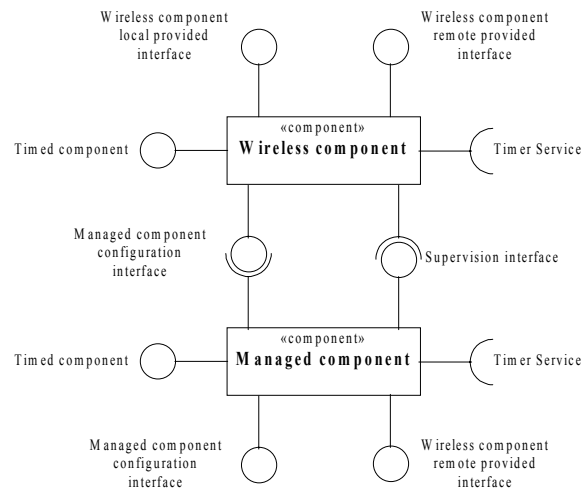


Figure 10. Detailed basic operations offered in each kind of interfaces from Figure 9

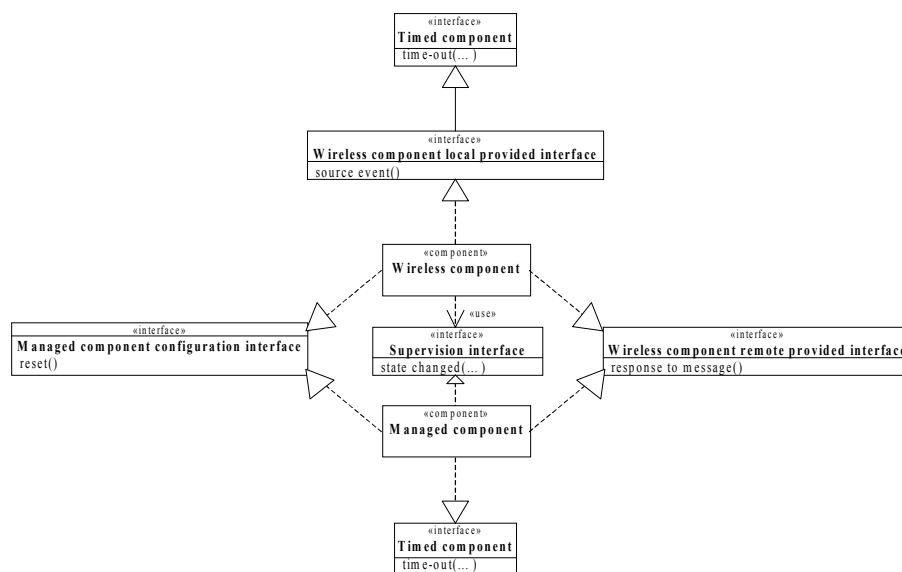


Figure 11. Alternative to Figure 10

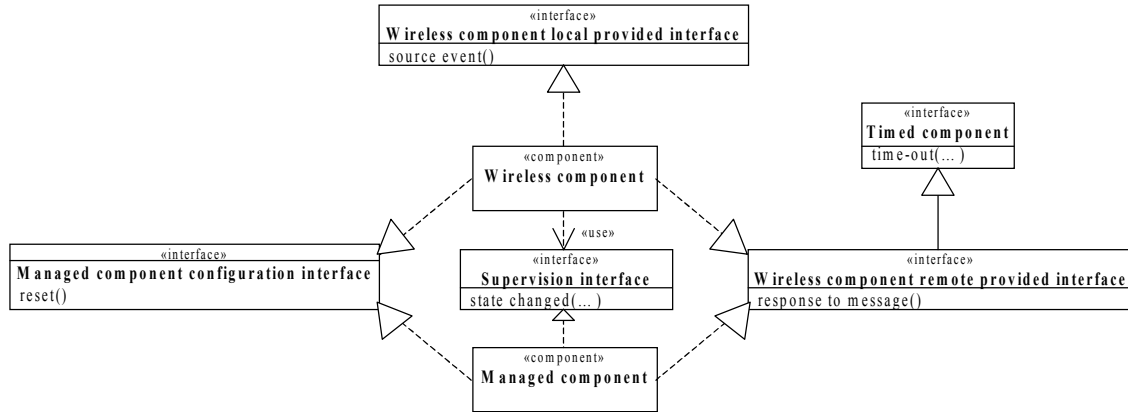
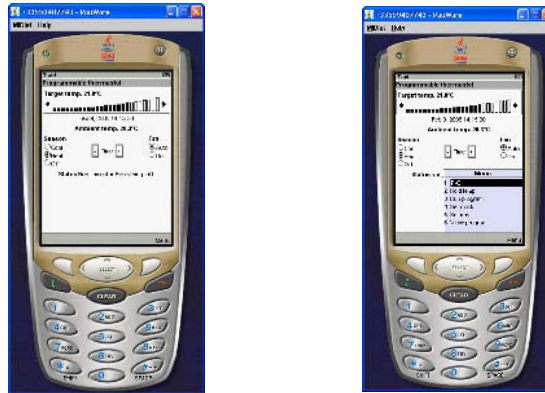


Figure 12. User interface of the home automation device and its possible evolution



sage that is interpreted in both automata in Figure 8).

In the administration context, management operations are assigned to special interfaces, namely *Managed component configuration interface* and *Supervision interface* in Figure 9. Finally, specific incoming events and flows for wireless components may lead to creation of similar phenomena on the managed components' side without any communication. In Figures 9 and 10, we take an example about timer services that are located and thus acquired

from different running platforms. This means here that interpreted *time-out* events on the wireless components' side are simulated with the same contexts and constraints on the managed components' side. More generally, this leads to the detailed and competing views in Figures 10 and 11 that raise the problem of control flow segmentation.

In Figure 11, *Timed component* is inherited by *Wireless component remote provided interface* instead of being independently implemented (symbol is white triangle with dotted

Figure 13. Component architecture of the home automation system

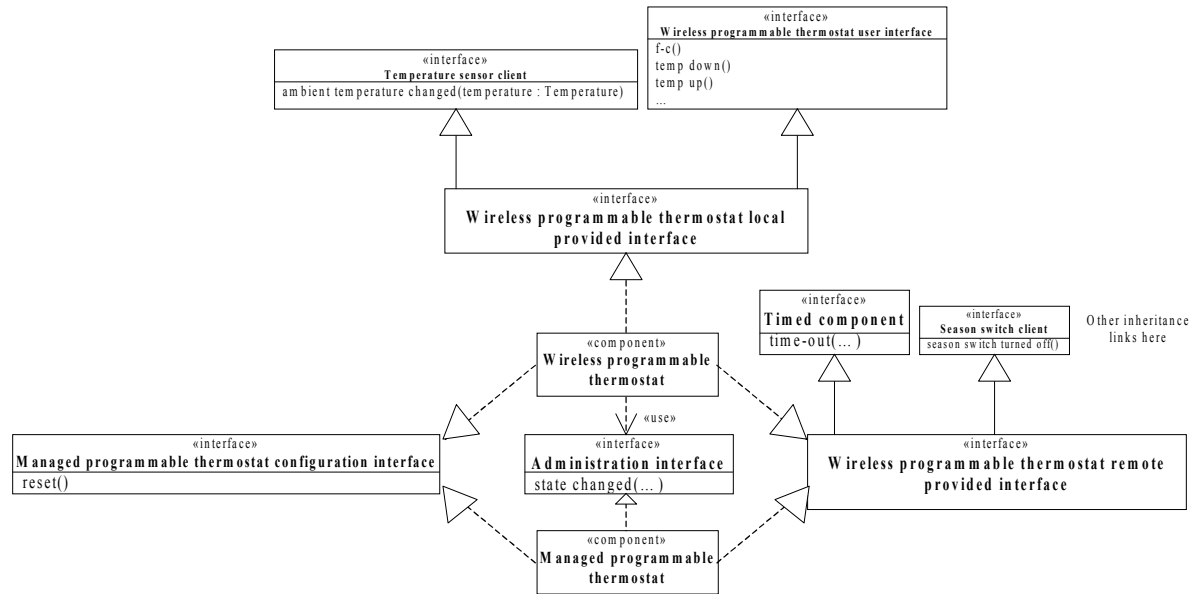


Figure 14. Web browser screenshot of the management environment



line in UML) by both wireless and managed components. Note that Figure 9 deliberately makes no assumption on this point. In Figure 11, being thus “included” in *Wireless component remote provided interface*, the *Timed component* interface makes managed components receiving *time-out* events from wireless components instead of using platform-dependent services (i.e., J2SE/J2EE services). So, both implementations of *Timed component* continue to differ, but that of managed components is fixed and unique, accepting *time-out* events coming from wireless components.

CASE STUDY: A HOME AUTOMATION SYSTEM

Figure 12 is the user interface that is an entry for some of the received events (e.g., *temp down*, *temp up*, *f-c*) of the complex statechart appearing in the Appendix. Although being complex, the UML statechart in the Appendix is easily and straightforwardly implemented thanks to the *PauWare.Statecharts* library (see the following code).

Figure 13 shows the chosen component architecture, namely the separation between what is locally received by the *Wireless programmable thermostat* component which implements *Wireless programmable thermostat local provided interface*, and what is received by the *Managed programmable thermostat* component through *Wireless programmable thermostat remote provided interface*.

For instance, temperature sensor events (*ambient temperature changed*) are acquired by *Wireless programmable thermostat* and propagate to *Managed programmable thermostat*, while switching events, such as *season switch turned off* which definitively stops the overall home automation system (see statechart in Appendix), are connected with any port of

Managed programmable thermostat and embody control commands when delegated to *Wireless programmable thermostat*.

Figure 14 is the final result showing all of the possible operations supported by the *Managed programmable thermostat* component in a Web browser. Visible buttons in windows of Figure 14 simply and straightforwardly map to service implementation for *Managed programmable thermostat*. Here is some Java code illustrating how the *PauWare.Statecharts* library helps implementation:

```
public void f_c() throws Statechart_exception {
    // f-c event (see Appendix and/or popup
    // menu in Figure 12)
    _BIT_programmable_thermostat.fires(
        _Ambient_temperature_displaying,_Ambient
        _temperature_displaying);
    _BIT_programmable_thermostat.fires(
        _Target_temperature_displaying,_Target
        _temperature_displaying,true,this,"switch_mode");
    _BIT_programmable_thermostat.fires(
        _Program_target_temperature
        _refreshing,_Program_target_temperature_
        refreshing,true,this,"switch_mode");
    _BIT_programmable_thermostat.run_to
    _completion("f-c");
}
```

The *run_to_completion* predefined routine conforms to UML 2 executability rules and, consequently, lasts through stable and safe automaton contexts. Management operations may thus occur as the *reset* user-defined function that may look like:

```
public void reset() throws Statechart
_exception { // configuration service
    to_state("Operate"); // Operate may
    be seen within the statechart in Appendix
}
```

CONCLUSION AND FUTURE DIRECTIONS

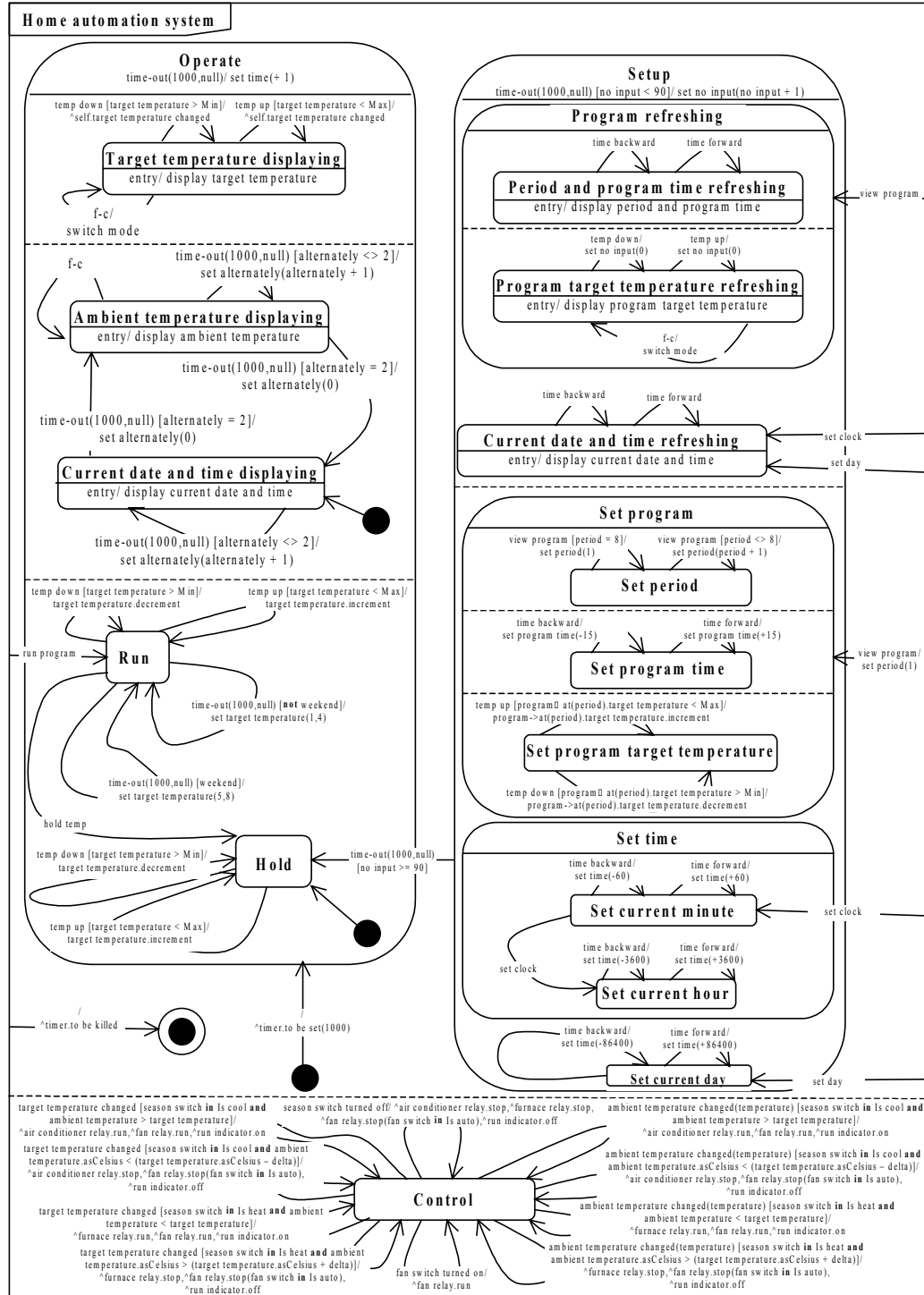
This chapter highlights the great need for administration of mobile environments, including the management of the collaboration between software components deployed in these environments and software components deployed in server-side tiers. Numerous hardware/software barriers preclude having full administration capabilities on wireless sides only. Furthermore, even in traditional management environments, logic and rationale for expressing supervision and control operations do not exist. For instance, JMX supplies a standard infrastructure but does not explain how to design manageable components.

The chapter shows that executable modeling, supported by the statecharts' reputable formalism associated with a dedicated library, both allow the methodical design of components for administration. The proposed approach is illustrated by a programmable thermostat wireless component.

REFERENCES

- Barbier, F. (2005). COTS component testing through built-in test. In S. Beydeda & V. Gruhn (Eds.), *Testing commercial-off-the-shelf components and systems* (p. 55-70). Berlin: Springer-Verlag.
- Bloch, C., & Wagner, A. (2003). *MIDP 2.0 style guide for the Java 2 platform* (micro ed.). San Francisco: Addison-Wesley.
- Bock, C. (2004). UML 2 composition model. *Journal of Object Technology*, 3(10), 47-73.
- Harel, D. (1987). Statecharts: A visual formalism for complex systems. *Science of Computer Programming*, 8, 231-274.
- Horrocks, I. (1999). *Constructing the user interface with statecharts*. San Francisco: Addison-Wesley.
- Kreger, H., Harold, W., & Williamson, L. (2003). *Java and JMX—building manageable systems*. San Francisco: Addison-Wesley.
- Mellor, S., & Balcer, S. (2002). *Executable UML—a foundation for model-driven architecture*. San Francisco: Addison-Wesley.
- National Coordination Office for Information Technology Research and Development. (2001). *High confidence software and systems research needs*. Arlington, VA: NCO/ITRD.
- Object Management Group. (2003). *UML 2.0 infrastructure specification*. Needham, MA: OMG.
- Object Management Group (2003). *UML 2.0 superstructure specification*. Needham, MA: OMG.
- Sun Microsystems. (2002). *Java management extensions instrumentation and agent specification, v1.2*. Santa Clara, CA: Sun Microsystems.
- Szyperski, C., Gruntz, D., & Murer, S. (2002). *Component software—beyond object-oriented programming* (2nd ed.). San Francisco: Addison-Wesley.

APPENDIX: STATECHART OF THE PROGRAMMABLE THERMOSTAT WIRELESS COMPONENT



Chapter XVI

Mobile User Data Mining and Its Applications

John Goh

Monash University, Australia

David Taniar

Monash University, Australia

ABSTRACT

Mobile user data mining is the process of extracting interesting knowledge from data collected from mobile users through various data mining methodologies. As technology progresses, and the current status of mobile phone adoption being very high in developed nations, along with improvements on mobile phones with new capabilities, it represents a strategic place for mobile user data mining. With such advanced mobile devices, locations that mobile users visit, time of communications, parties of communications, description of surrounding locations of mobile users can be gathered, stored and delivered by the mobile user to a central location, in which it have the great potential application in industries such as marketing, retail and banking. This chapter provides a general introduction on mobile user data mining followed by their potential application. As the life of mobile users are mined, general patterns and knowledge such as the sequence of locations they tend to visit, groups of people that they tends to meet, and timing where they generally active can be gathered. This supports marketing, retail and banking systems through the use of knowledge of behavior of mobile users. However, challenges such as privacy and security are still a main issue before mobile user data mining can be implemented.

INTRODUCTION

With the increasing penetration rate of mobile technologies among the marketplace (Goh & Taniar, 2004b; Lim, Wang, Ong, & Hwang, 2003), businesses have adopted various types

of mobile products, such as personal digital assistants, mobile phones, and wireless laptop computers, in order to help improve the efficiency of one's daily life. The increasing adoption of such equipment allows the opportunity for the collection of data about their usage and

movement that can then be further analyzed (Goh & Taniar, 2004b, 2004d).

The analysis of such data collected from mobile devices and mobile users for determination of patterns is called *mobile user data mining* (Goh & Taniar, 2004a, 2004b, 2004c, 2004d, 2005). In the mobile environment, there are devices offering service to mobile equipments. These are often known as static devices (Goh & Taniar, 2004b, 2004d), as they stay static and provide services for the mobile devices. These mobile equipment operate in a network where data can be readily transferred and services can be readily rendered (Goh & Taniar, 2004c).

Data mining can be performed in various domains such as the time series domain (Han, Dong, & Yin, 1998, 1999; Han, Pei, & Yin, 2000), Web domain (Christophides, Karvounarakis, & Plexousakis, 2003; Dourish, 2004; Eirinaki & Vazirgaiannis, 2003; Kastaniotis, Zacharis, Panayiotopoulos, & Douligeris, 2004; Kim, Kim, & Kim, 2004), market-basket analysis domain (Agrawal & Srikant, 1994, 1995), geographical information system domain (Koperski & Han, 1995), performance improvement domain (Han et al., 2000; Li, Tang, & Cercone, 2004; Miyahara et al., 2004; Thiruvady & Webb, 2004; Yip, Wu, Ng, & Chan, 2004), security domain (Oliveira, Zaiane, & Saygin, 2004), and mobile domain (Goh & Taniar, 2004a, 2004b, 2004c, 2004d, 2005; Wang, Lim, & Hwang, 2003). The existing methods of data mining include association rules (Agrawal & Srikant, 1994) and sequential patterns (Agrawal & Srikant, 1995). The existing methods of mobile user data mining include frequency patterns (Goh & Taniar, 2004b), location-dependent mobile user data mining (Goh & Taniar, 2004d), and parallel patterns (Goh & Taniar, 2004c).

The aim of this chapter is to provide an insight on the background of mobile user data mining and potential application areas of mobile

user data mining in different industries. The potential application is viewed from the aspects of the banking industry, marketing industry, and retail industry. As the mobile user data mining methods are getting more developed, they could be implemented one day in areas where interactions are required with highly mobile customers.

This chapter is organized as follows. The next section provides a background on mobile user data mining. We then highlight the benefits of adopting mobile user data mining and provide detail about how banking, marketing, and the retail industry could benefit from mobile user data mining. Next we provide an overview of future challenges such as security and privacy, and finally summarize the chapter.

MOBILE USER DATA MINING

Mobile environment (Goh & Taniar, 2004a, 2004d) refers to an environment where free movements of human beings are possible. The mobile environment is an area where a human being carries devices that can be handheld (mobile devices), and seeks information and services from service providers (static devices) that is within the coverage of that area or available through subscription. Mobile environment therefore can also be referred to as the mobile network environment (Goh & Taniar, 2004d).

In the mobile environment, many devices are being used. The static devices are nodes in the mobile that never moves around, but are stationed in the mobile environment just to provide services such as data, network, and processing to their clients, which are mobile devices (Goh & Taniar, 2004b, 2004c, 2004d). These mobile devices are wirelessly connected together to a server. The mobile devices can come in many forms. Some of the existing forms include mobile phones, personal digital assistants, and laptop computers. Mobile de-

vices are the devices that follow or are being carried by mobile users (Goh & Taniar, 2004b). They generally have less processing, network, and data capacity, and need to request service from static devices to become useful. Mobile users are human beings that carry the mobile devices in the mobile environment for the purpose of finding out location-dependent information about a current locality (Goh & Taniar, 2004d), navigating using mobile devices (Lim et al., 2003), and communicating with other mobile users in the mobile environment (Goh & Taniar, 2005; Lim et al., 2003).

Mobile user data mining (Goh & Taniar, 2004a, 2004b, 2004c, 2004d, 2005) is the activity that uses data collected from mobile equipment for analysis. Mobile equipment in this context can include mobile phones, personal digital assistants, and laptop computers. Data mining allows large amounts of data to be analysed in order to find out interesting patterns about the data. Two examples were given:

1. The use of data mining techniques to find out whether consumers tend to visit location *A* immediately after location *B* (Goh & Taniar, 2004d, 2005).
2. The use of data mining techniques to find out whether consumers tend to buy product *A* and product *B* at the same time in one single shopping trip (Goh & Taniar, 2004d).

The prerequisite for mobile user data mining (Goh & Taniar, 2004a, 2004b, 2004c, 2004d, Mar 2005; Lim et al., 2003) to work is the availability of a large amount of dataset collected from the mobile user (Goh & Taniar, 2004b). This prerequisite should be achieved as much as possible. The increasing rate of adoption of mobile phones by the world population to use as a main method for telecommunication helps achieve this prerequisite. The wide adoption of mobile phones provides the opportunity for the data to be gathered for mobile user data

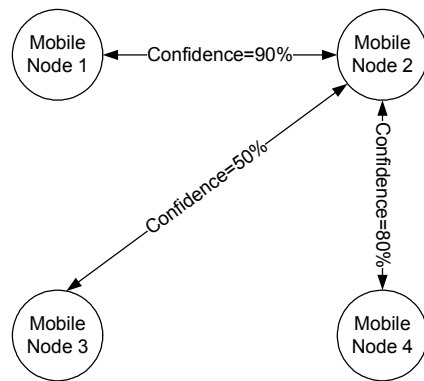
mining (Goh & Taniar, 2004a, 2004d). By obtaining these data, the following are possible to gain more insight about the mobile users:

- a. **Hypothesis testing**, such as testing the data whether mobile users in the population visit location *A* immediately after location *B*.
- b. **Data analysis**, such as performing statistical functions in order to report on the statistical report on the population and categorize the population into different segments for other purposes such as target marketing.
- c. **Knowledge generation**, such as applying data mining methods on the vast amount of data gathered to generate a list of rules and patterns about mobile users. Rules and patterns are then refined into knowledge.

Mobile user data mining (Goh & Taniar, 2004a, 2004b, 2004c, 2004d, 2005) involves the identification of a mobile device and mobile user, and then joining the data to represent the activity of the mobile user. This data is then used as the source data to be fed into mobile user data mining algorithms in order to find out the underlying patterns on the data. The identification of a mobile device (Goh & Taniar, 2004b, 2004d) might include a mobile phone number, mobile phone model, mobile phone university identification number, owner of a mobile phone, current date and time at the place of a mobile phone, current location of a mobile phone, current context of a mobile phone, current entity the mobile phone is communicating with, current mobile phone power, and memory status.

Identification of a mobile user (Goh & Taniar, 2004a, 2004c) involves variables such as user name, address, e-mail, occupation, age, educational background, marital status, family members, personal interest, business interest, cur-

Figure 1. Frequency pattern



rent activity, time of peak activity, race, nationality, transportation vehicle detail, criminal history, and religion.

Some of the variables are less restricted. Some variables that are more sensitive require more consideration and preparation in the future so that data can be gathered for the benefit of all parties (Goh & Taniar, 2004b). Mobile user data mining (Goh & Taniar, 2004b, 2004d) can thus be summarized as data mining with the use of variables of personal details of mobile users and the current environment details where the mobile user is currently physically located, in addition to the ones normally utilized, such as items bought.

Mobile User Data Mining Techniques

Mobile user data mining aims to analyze and predict behavior of mobile users from the data collected by activities generated from mobile users. Some existing mobile user data mining methods are given below.

Frequency Pattern

Frequency pattern (Goh & Taniar, 2004b) uses the frequency of communication with pre-specified

criteria in order to find out the logical proximity of mobile users. In a mobile environment, it is often the case that mobile users are using the mobile equipment to communicate with another mobile user. This can be done either by voice, voicemail, text, and so forth. Frequency pattern takes account of all these communications and records them into a list. There is a time series which encompasses the time desired immediately from a current time period. Different phases of the time series are given different weight in terms of the relative importance to be significant as frequent communication.

Figure 1 illustrates frequency pattern (Goh & Taniar, 2004b). Circles labelled 1, 2, 3, and 4 are the set of mobile nodes that is found to have close relationships, as they are all joined together by using arrows that show high confidence (*more than 50%*). The relationships between the mobile nodes are determined by finding out their frequency of communication without considering their physical distance among each other, and passing it to a pre-specified criteria which determines which part of the time window is more important. The final output of frequency pattern is a set of mobile nodes connected by relationships (*represented in arrows*) which suggests close relationships among them.

The generation of time series with the frequency of communication along with the relative weightage is called the pre-specified criteria. Overall, the higher emphasis is usually placed nearer to the current time point. The frequency of communication along with the pre-specified criteria will determine the confidence of frequency of communication in between two mobile users. When two mobile users have been determined as communicating frequently, they will be marked as a frequent relationship in between them.

The list of frequent relationships within a mobile environment tells the possibility that the

list of mobile users connected by frequent communication may represent a group that is closely related. For instance, they could be a group of close friends or close family members that frequently stay in touch with each other. The limitation of frequency pattern comes from the fact that it is expensive to mine out the frequency patterns when mobile users come in a large volume. Frequency pattern requires the finding of frequency of communication from itself with every single other mobile nodes during each mobile user data mining exercise. This causes the requirement for a high amount of overhead data, which reduces the performance of data mining. Instead of mining by gathering all data, we gather data from the mobile node itself.

Location-Dependent Mobile User Data Mining

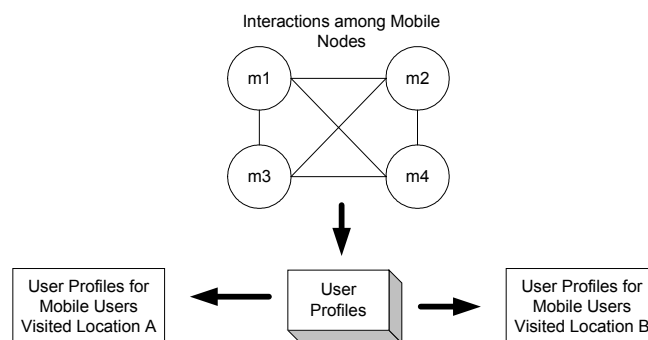
Location-dependent mobile user data mining (Goh & Tanian, 2004d) is a method that uses user profiles in order to find out knowledge specific to a particular physical location, using previously available data mining algorithms, such as association rules and sequential pattern data mining. It is often useful to find out knowledge specific to a particular location, such as during instances when a decision that relates to a particular physical location has to be made.

For example, a manager of a store wishes to find out the behaviour of mobile users at large, but who have visited the store of the manager.

User profile (Goh & Tanian, 2004d) is a profile which stores information about a particular mobile user. The knowledge of user profile is organised in the format of *mobile_user_identification* ($theme_1 = x\%$, $theme_2 = y\%$, ..., $theme_n = z\%$). The user profile is built up over a period of time by engaging regular communication with static nodes that are constantly serving and recording the resources utilised by the mobile nodes. For example, over time, mobile nodes visited various physical locations. Each physical location is associated with a certain theme, such as *shopping*, *entertainment*, *academic*. These themes are regularly updated to the user profiles; the more occurrences of such events, the higher the confidence value will be associated with a particular theme.

Once this process have been occurring over a period of time, the user profiles will be more accurate, as more transactions are fed into the system and the value of the user profiles better reflects the knowledge gathered about the mobile user over a long period of time. Once this has been achieved, then the mobile user data mining process can be started. This is achieved by first identifying the list of mobile users that have visited a particular physical

Figure 2. Location-dependent mobile user data mining

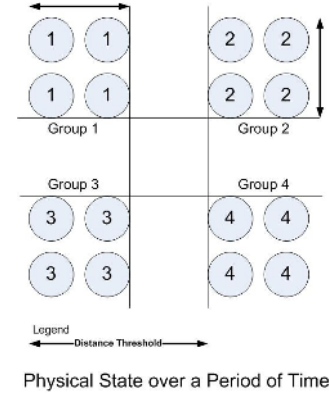


location over a period of time. This serves as a log of who has been visiting the physical location. The process then matches the mobile users identified from the list by querying the information collected about them—that is, the user profiles form a temporary database. This temporary database contains the list of mobile users who visited a physical location, and their visiting behaviours that are drawn from user profiles. Once the matching is done, then it can be passed to any other algorithms in order to find different types of knowledge from the system.

Figure 2 illustrates the location-dependent mobile user data mining (Goh & Tanir, 2004d) scenario. The interactions among mobile nodes, which are represented as m_1 , m_2 , m_3 , and m_4 in circles are collected and recorded and passed to user profiles. Each time the source data is passed to user profiles, the database is updated by counting the number of visits the mobile nodes have done, and the list eventually becomes more accurate to represent the behaviour of mobile users. Once this is done, the transactions in user profiles are queried by a location-dependent miner, by first identifying all the mobile users that visit the physical location. After this, the list gathered for a particular location can be passed to other data mining algorithms such as association rules or sequential patterns.

Location-dependent mobile user data mining (Goh & Tanir, 2004d) has the weakness at the user profile side, in which the accuracy of the user profiles determines the accuracy of the final result. The building of user profiles takes time and coordination among static nodes. It places the requirement at the static node to identify and record incoming and outgoing mobile nodes. Another weakness of location-dependent mobile user data mining draws from the heavy dependence on the user profiles. Issues arise when the mobile nodes that visited a physical location have not been registered in

Figure 3. Group pattern



the user profiles at all, and this will distort the final result of the data mining system. Parallel pattern will solve this method by using the similarities of decisions in order to perform data mining instead of using user profiles. The proposed method further enhances the parallel pattern by examining the relationships among these parallel patterns.

Group Pattern

Group pattern (Wang et al., 2003) is a mobile user data mining method which can be used to find useful knowledge from source data collected from mobile users. Group pattern (Wang et al., 2003) analyses the physical distance in between mobile users over a time series and determines the group relationships among mobile users. The physical distance used was Euclidean distance. A list of mobile users are qualified to be a group pattern when there has been at least one occurrence when all of these mobile users have been physically close to each other ($distance < distance\_threshold$) and over a period of time ($duration > duration\_threshold$).

Figure 3 shows a graphical representation of four group patterns (Wang et al., 2003). Group pattern (Wang et al., 2003) is a physical distance-dependent mobile user data mining method, in which mobile users who stays close

enough together with each other over a certain time threshold will be qualified as one group pattern. The four groups are distinct from each other: since it is physical-distance based, one mobile user can only belong to one group pattern (Wang et al., 2003), and not another, because one mobile node cannot be present at two locations at one time, and thus cannot be a member of two group patterns.

The advantage of using physical distance in order to perform mobile user data mining is that in a confined environment, and when there is no communication occurring among mobile users, this method is useful. However, during situations when the geographical area of the mobile users in concern are large and there are many communications which do not need the mobile users to be physically close, frequency pattern will be useful, as frequency pattern (Goh & Taniar, 2004b) does not take the physical distance but frequency of communication into consideration when finding the group relationships of mobile users.

Benefits of Mobile User Data Mining

The main advantage of possessing mobile user data mining technology is the ability to deliver personalized services to customers (Goh & Taniar, 2004b, 2004d). The power to personalize comes from the ability to understand individual needs out of a mass amount of customers by the assistance of data warehouses and data mining methods.

In order to achieve personalization, the understanding of what the customer does in the whole day, everyday, has to be used as the input for knowledge generation. Mobile phones were used primarily for mobile user data mining to gather information about mobile users (Goh & Taniar, 2004b, 2004d). The fact that mobile phones were carried by mobile users all the time fulfills the condition of mandatory need for

gathering data regarding day-to-day activity of the customer.

The ability to gather and process such complete data that presents the daily life picture of customers enables the generation of precise knowledge about mobile users (Goh & Taniar, 2004c, 2005). The latest mobile phone technology allows the mobile user to be identified. The problem of one mobile phone used by different mobile users is then eliminated.

Cost benefit of mobile user data mining can be measured by means of key performance indicators. The cost can be calculated as the cost for generation of knowledge per customer. The cost per customer varies depending on (Goh & Taniar, 2004b):

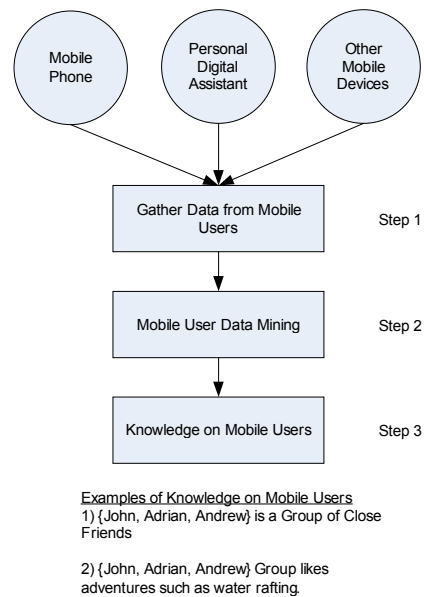
- a. the data mining algorithm used,
- b. the size of customer records, and
- c. the pre-existing summarized knowledge about the customer.

In order for a mobile user data mining project to be successfully implemented, the cost of finding knowledge per customer should be lower than the benefits for knowing what the customer wants. The main criterion is that the customer must be a customer that, if the personalized service is provided to them, will return business transactions, which increases business earning.

The cost of finding what the customer wants remains high (Goh & Taniar, 2004b). This is due to the high cost of searching for information about the customers. The searching for information about the customers without the use of technology will be prohibitively expensive. The current method of data gathering is limited to the point of sales data gathering.

Point-of-sale data gathering involves recording what the customer buys in that shopping center and saves it according to a loyalty program customer identification number. Mo-

Figure 4. Mobile user data mining as an analysis tool for the marketing industry



Mobile user data mining, through the use of a mobile phone, enables the ability to gather individual details of customers in a cost-effective way. Therefore, mobile user data mining will continue to exist.

POTENTIAL APPLICATIONS

Potential application of mobile user data mining is most promising in the area where frequent interaction with customers that are highly mobile is a norm. Here, three industries—namely marketing, banking, and retail—are selected in order to provide an insight of the potential applications of mobile user data mining techniques in these areas. Mobile user data mining can value add the marketing industry through the gathering of data from mobile sources about the market population in real time for market research. The banking industry deals with many transactions, including mobile transactions, with a wide range of variety of customers that travel nationally and internationally, and also requires

a high degree of data integrity and confidentiality. The retail industry interacts with different customers, who can be identified through their mobile location-dependent and time-dependent variables. Prediction of their behavioural patterns can be found using these as a basis for analysis.

Application in Marketing

Marketing is a suitable implementation ground for mobile user data mining. Mobile user data mining value adds marketing through the gathering and analysis of mobile data that provides intensive customer information. The result is better marketing decisions, through better quality of information.

Marketing environment involves market research to perform research on data gathered on the market in order to find out trends and behaviors of consumers by means of various methods, including statistics and data mining. Marketing environment also involves supply of market knowledge. The analysis of the market

is then documented, and reports were supplied to subscribers from different lines of business from different industries, to provide an understanding of what the market is and what needs to be done.

Mobile user data mining provides assistance for the analysis of market data. The market analysis using mobile user data mining is a large-scale analysis in the marketing industry. Mobile user data mining, through the analysis of movements of mobile users, and which includes the time and location of the mobile users, supports the ability to identify and analyze the mobile market population.

Figure 4 illustrates where mobile user data mining fits in to the process of building a consumer knowledge base for marketing business. A consumer knowledge base is a base of knowledge in which knowledge about current trends and patterns are stored. A business in the marketing industry gathers data from various sources, including consumer survey, mobile phone, sales data, population, and qualitative analysis. The data mining processes include the traditional data mining techniques and also the mobile user data mining techniques.

Data gathered are passed through a process which contains a list of tools. These analysis tools are statistical tools and data mining tools. Mobile user data mining, in particular, focuses on the processing of data gathered from mobile phones. The resulting knowledge generated is then contributed towards the consumer knowledge base. It is a valuable intangible asset for the business for later sales to other businesses in different industries.

The marketing industry collects data from different segments of populations. This is done by many means, including telephone interview and survey. Such collection method will have the potential tendency to collect inaccurate data. Inaccurate data can be caused either by incorrect recording through asking the wrong question, or the consumer purposely releases inaccurate information.

One approach to solve this problem is through the use of database matching, whereby individual records are checked with other sources to ensure that the consumer has released the same information every time.

```
Function Consumer Provides (Marketing Answers) {
    Checking for Missing Values (Marketing_Answers);
    Checking for Invalid Values (Marketing_Answers);
    Check Consumer History (Consumer.Marketing_Answers);
    If Record on Consumer Found Then {
        Check Validity (Join (Marketing_Answers, Consumer_Record));
        If Valid Then Save ();
        Else If Invalid Then {
            Mark As Invalid ();
            Discard ();
        }
    }
    Else {
        Mark Record as Uncertain ();
        Save Consumer Record ();
    }
}
```

The second approach for resolving such an issue is through the use of more reliable channels of survey, such as using mobile phone tracking to gather consumer behavior data.

The third approach for resolving such an issue is through the use of data mining methods which use predictive analysis to determine the confidence of data. This concept works by means of using data mining to make predictions; for example, if the consumer aged between 20 to 30 and visits a sports and recreational facility very often, the consumer is more likely to buy x y z items in the supermarket.

```
Function Consumer Provides (Marketing Answer) {
    Prediction = Data Mine (Consumer_Background);
    Match (Prediction, Marketing_Answer);
    If Match () = True {
        Mark Consumer as Normal Category ();
    }
    Else If Match () = False {
        Mark Consumer as Special Category Requiring Human Check ();
    }
}
```

Protection of Derived Knowledge

During the operation of performing data mining, different entities may require different levels of cooperation by means of sharing of data col-

lected and sharing of knowledge generated from the data collected. As the cost for generation of such knowledge is high, and the cost of losing the data to competitors will cost the entities, it is important to safeguard the information through security means.

An example is when derived knowledge such as Group {John, Adrian, Andrew} is a group of close friends who enjoy sailing. This was found because of the data provided from the shop selling sailing equipment and data provided by the telecommunication service. This piece of information is valuable for anyone who wishes to do business with this group. There were costs incurred to gain such knowledge, and the protection of such knowledge through limiting sharing is essential. For instance, a business can buy knowledge about John, but will not gain access to the knowledge related to the whole group.

The use of security features is the existing method for resolving these issues. Existing security methods include the use of data protection, such as passwords, and public and private keys, to ensure data confidentiality, integrity, and availability.

However, these were not enough in the current world where information sharing is done through real-time sharing over a network which can span across the globe. The protection of such data may hinder the performance of information sharing.

In order to resolve this issue, the data mining field has previously proposed concepts of secure sharing of data during data mining.

```
Function Data Sharing () {  
    Define Attributes to Share ();  
    Define Individual Levels of each Attribute Sharing ();  
    Transfer Sharable Attributes to Common Sharing Zone ();  
    Limit the Duration of Sharing ();  
}
```

Samples of outcome derived from mobile sources through mobile user data mining methods that are applicable to marketing industry

are provided. Mobile user data mining contributes to marketing industry through location-dependent, time-dependent, and communication-dependent knowledge previously not feasible to gather. The following knowledge can then be updated to individual profiles in order to enhance the marketing knowledge. Techniques applicable are: frequency pattern, parallel pattern, and location-dependent mobile user data mining.

Frequency Pattern: {User1, User2, User3} is a group of close friends.

Parallel Pattern: {User1, User2, User3} tends to move from Sports to Cinema together.

Location Dependency: Sports location is frequented by people who like {adventure sports, golf, and fishing}.

Marketing Profile User1: adds {Sports, Cinema, closeTo (User2, User3)}
Marketing Profile User2: adds {Sports, Cinema, closeTo (User1, User3)}
Marketing Profile User3: adds {Sports, Cinema, closeTo (User1, User2)}

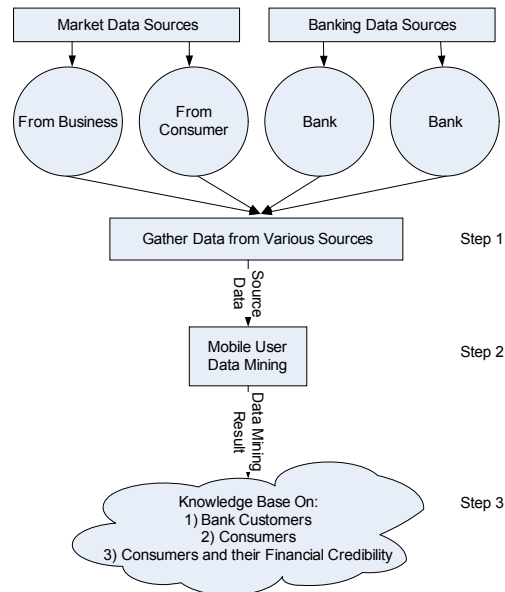
Application in Banking Industry

The banking industry involves the activity of gathering financial resources from the population and businesses, and redistributing the financial resources to other entities needing them, benefiting by acting as a financial intermediary. Banking used to be brick-and-mortar based, where consumers have to go to a physical banking outlet to do banking. With the increasing use of information technology, banking systems have developed to be accessible by means of the Internet and mobile devices.

The banking industry can capitalize on mobile user data mining by using its strategic position as a financial intermediary. Consumers' behavior can be represented strongly, simply by viewing what the consumer buys. Consumers have a limited amount of financial resources, and the decision to buy something will mean that it is something important—that other products or services are passed over in order to take up this product or service.

The increasing use of mobile commerce is most visible in the example of being able to

Figure 5. Mobile user data mining as a fraud analysis and prevention tool in the banking industry



purchase meals or buy a soft drink from a vending machine simply by pressing buttons on the mobile phone, deducting the credit from the mobile phone account. With the increasing use of electronic means and mobile means for transactions, these can be gathered easily by the bank, which stores the financial resources of the customers.

The bank is in the best position to analyse the behavior of a customer based on the monetary view. The monetary data of the customer is impossible to be gathered by other entities besides the bank itself. Moreover, almost all people and businesses in the population will had to have at least one account with a financial institution. The financial data can be utilized to perform knowledge generation, while preserving the privacy and security of financial data.

The key advantages include that knowledge generated can be used to benefit the operation of the bank by providing the most relevant

financial services to the customer. The knowledge could be used as an asset for services that provide knowledge about the customer, such as a consumer group's buying behavior to other business entities while preserving privacy and security. The knowledge can be used to join with other official entities, including the government and other financial institutions, to form an overall picture of the customer to minimize fraud. The technology could facilitate the gathering, recording, and sharing of customer credit knowledge.

Figure 5 illustrates mobile user data mining as a fraud analysis tool and a data analysis tool in the context of a bank. The bank has a strategic position to have the authority to gather and analyze the financial data of customers. The transaction data are gathered from business entities and consumers, all from the market. Data mining, including mobile user data mining, is then performed to produce knowledge on customers. This is stored in the knowledge base. The knowledge base could further be used for:

- legal enforcement for tracking fraudulent transactions,
- sale to the service industry on consumer behavior in a specific industry while preserving privacy and security,
- sale to the marketing industry to better analyze the population behavior, and
- sharing among financial institutions for the prevention of fraud and credit rating checks.

When confidential financial data of customers are shared among banks, the integrity of those data must be preserved. The bank has the duty of care to ensure that financial data are absolutely accurate, if they wish to participate in data mining. This is an issue when data are shared and processed by data mining software.

The existing method for preserving integrity of data is achieved through versioning control and backup storage of data.

Customer Data (Name, Address, Balance, Check Digit) – Stored in Location A
Customer Data (Name, Address, Balance, Check Digit) – Stored in Location B

There is a need for development of methods built into the data mining methods which could preserve the integrity of customer financial data with minimal amount of data mining performance issues.

Privacy of Customer Identity

When data mining is performed, privacy of customer identity must be preserved to ensure that customers cannot be identified if the customer records are shared outside of the organization. This is a research challenge, because it involves the need to shadow the customer identity, but still provide the ability to identify their unique identity and reverse this identity back to the original identity.

Existing methods include the use of mapping, which converts the customer identity to another value, which then provides identification on the other end.

Customer A (In the Organization)	Customer A (Out of Organization)
John	Q23
Mark	F21
Susan	U34
Jane	P23

The map needs to be recorded, and conversions are required to reverse the display of customer identity. The conversion of such customer identity includes Q23, U34, which provides an unknown pattern so that outside organization will not be able to perform the reversal to determine the customer. In the following case, it displays a case where conversion out of the organization without the map is possible.

Customer A (In the Organization)	Customer A (Out of Organization)
John	M1
Mark	M2
Susan	F1
Jane	F2

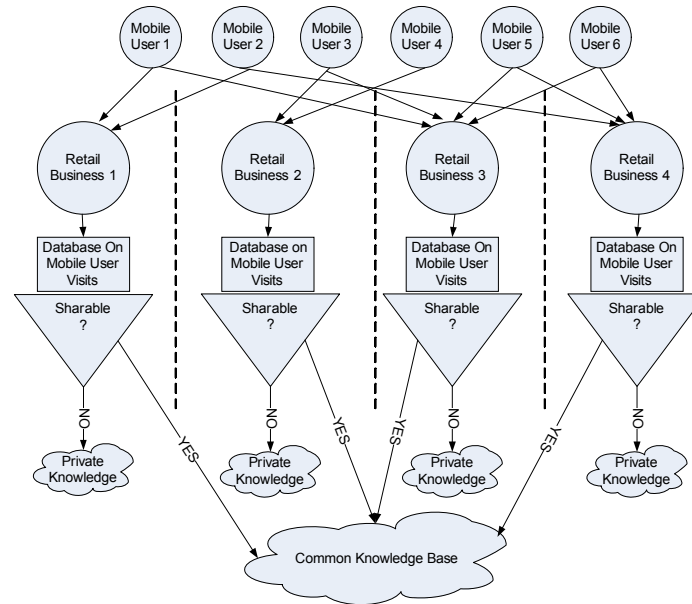
The resulting knowledge view out of the organization can be as follows: {Q23, F21, U34} is a group of close friends. They enjoy products related to fishing. If a retailer wishes to target this group of customers, a retailer could subscribe to a service or pay per reveal, such as it will cost \$100 to find out where the suburbs Q23, F21, and U34 are. It will cost another \$100 to reveal their complete mailing addresses.

The pattern of the above is clear: M represents Male, and F represents female. Such conversion will lead to easy reversal, which will affect the privacy issues of the originating organization.

Samples of outcome derived from mobile sources through mobile user data mining methods that are applicable to the banking industry are provided. Mobile user data mining contributes to the banking industry through the data collected from mobile devices as a result of mobile commerce. In addition, location-dependent information and time-dependent information are also collected so that the movement activities of mobile user and the commercial activities of mobile users are used for mobile user data mining. Techniques applicable are: frequency pattern, parallel pattern, and location-dependent mobile user data mining.

Case 1: Banks Location Dependency: Finds out fraud-prone or crime-prone areas. Resources are then directed on these areas for a higher level of security checks.
Case 2: Banks Reported that {User1} has fraudulent record or intention.
Frequency Pattern: Investigates the fraud rings, through the frequency and timing of communications. {User1, User4, User7} often communicate with each other, thus suggesting some working relationships among them.
Parallel Pattern: {User1: Restaurant – Bookstore}, suggesting that the fraudulent user may be traced at the bookstore, if it cannot be found in the restaurant.

Figure 6. Mobile user data mining as a knowledge sharing tool in the retail industry



Application in Retail Industry

The retail industry involves a wide range of businesses selling goods and services to consumers. The businesses in the retail industry are the transaction points where customer transactions take place. The gathering of knowledge in the retail industry works by:

- the wide range of goods and services sold to consumers which is used to identify the taste and needs of the consumer;
- the heavy use of electronic transaction means, including mobile commerce, card purchases, and online purchases; and
- the large network of businesses in the retail industry willing to share their information among each other, while preserving privacy and security.

The retail industry is the largest industry in the discussion of this chapter in terms of the number of businesses and their geographical coverage. The retail industry can capitalize on

mobile user data mining by the existing architecture that is widely covered and densely networked. This is a strength whereby knowledge can be derived by sharing the information with privacy and security among other retailers, and is less reliant on the marketing industry to perform the analysis of market.

The retail industry style of mobile user data mining uses a distribution strategy. By dividing the task of collecting consumer data to individual businesses in the wide area of the retail industry, these data are then shared on a common knowledge base. Retailers who are on the network could then subscribe to the knowledge base, and gather the market information without the compromise of privacy and security. This knowledge base can also be further shared with other business entities who wish to pay for such service.

Figure 6 illustrates the model in which mobile user data mining is used as a distributed knowledge sharing tool in the retail industry. The retailers have a widely covered and densely networked system where transactions with the

consumers are recorded and are shared among other retailers subscribed within the same network. Each will perform its own data analysis and data mining function, including the use of mobile user data mining that utilizes mobile phone data in addition to generic data. The outcome of the analysis is stored in a knowledge base which is shared among subscribers of the network, including entities outside of the retail industry, such as the marketing industry and the banking industry.

Samples of outcome derived from mobile sources through mobile user data mining methods that are applicable to retail industry are provided. Mobile user data mining contributes to the retail industry through data collection from mobile devices from mobile users, and the sharing of such individual knowledge among all other retailers through a central service provider so that the knowledge can be well utilized. Techniques applicable are: frequency pattern, parallel pattern, and location-dependent mobile user data mining.

Knowledge of Retailers
Retailer 1: Knows only about User1 {likes cappuccino}.
Retailer 2: Knows only about User2 {likes blue colored basketballs}.
Retailer 3: Would like to know about User1, User2, and User3.
Mining Performed by Service Provider
Frequency Pattern: {User1, User2, User3} are close friends.
Physical Parallel Pattern: {User1, User2} often goes from Retailer 2 after Retailer 1.
Service Provider Facilitates Knowledge Sharing
Location-Dependent Mobile User Data Mining: Location of Retailers 1 & 2 shows that people who visit there tends to like {newspaper, breakfast}.
Retailers 1, 2: Then sells newspaper or complimentary breakfast to attract customers.

FUTURE CHALLENGES

Mobile user data mining is indeed a valuable tool worth implementation to assist in finding relevant knowledge about the customers and potential customers. The future remains bright; however, there are some issues that need to be addressed in the future. Selected future challenges include: privacy issue on data handling, protection of data (data security) issue, user identification issue, and handling of massive volume of real-time data issue.

Privacy Issue

The privacy issue remains an important element that has been widely implemented in developed countries. The privacy issue exists mainly due to the need for respecting human rights, being able to do things freely with minimal interference from others. This includes the ability to remain anonymous and unknown of their choice and behavior, unless required by a legal or political system.

Personal details are being used extensively in mobile user data mining in order to produce a rich dataset. Privacy laws will indeed create barriers to access to such information. This issue could potentially be addressed by means of privacy-preserving methods, such as by a limited amount of view of the personal information through summarized data, or anonymous data which only identifies a person with a primary key of just numbers. While preserving the privacy of customers, the other factor of maintaining data consistency is also important so that the result of the analysis generates accurate knowledge for later use.

For example, data gathered contains financial data, location visited data, personal information data, and is shared among different entities from banks, retailers, and market research firms. Not all firms should be allowed to have full view of all data. All of the entities should not be able to reverse the output back to input to gather sensitive information. A technique that divides the data into financial, market research, and personal information that confidentially transfers to the authorized entity when required, in an accurate manner, is the desired result.

Data Protection

Protection of privacy of customer data involves protection of such data so not to be accessed by unauthorized parties. Protection from duplica-

tion or sharing is the key aspect of data protection here, as this knowledge has been gathered through investment and is an asset to the business. These assets are valuable for business in general, as the ability to have such information provides better decision making, creating competitive advantage. Due to the existence of knowledge in electronic form, duplication and sharing is likely. A strategy coupled with technology and legal solution is required to preserve the value of the asset.

For example, protection of data generated is an essential tool. Consider a business, having invested in the gathering of data through a high amount of effort, finding useful patterns from the data. The firm could provide service for other firms, while retaining its benefit of possessing the data through investment by preventing other firms from sharing the data.

User Identification Issues

As mobile user data mining will involve the identification of many different users in the mobile environment, errors and confusions might occur. Identification has many different purposes, including distinguishing one from another to ensure that for a series of identification numbers 1 to infinity, each represents a different human being. The identification of a user is based on the purpose and requirement of the technique used. If a technique requires the finding of similar activities performed by a single human being, then the identification of the person in question is required, such as through a national identification system.

If the technique focuses the items bought and location visited, then the identification requirement is only the identification of the mobile device. This can be done through the use of a network identification number, which is a unique number for every single network hardware that can be used to connect to a network. The

selection of the proper identification strategy and technique will be based on the identification requirement of the purpose of mobile user data mining.

For example, consider that a group of mobile users needs to be identified for tracking their financial records. The mobile user can be identified individually through the use of username and password, or credit card details, as only one individual can logically be associated with the identification means. On the other hand, when identifying users that have visited a mall and the place that they have visited soon after, it is only required to identify the mobile device through a number, such as the serial number of a network device identification number such as 00-F0-30-4E-4A.

Handling of Massive Volume of Data

When operationally active, mobile user data mining often will handle a massive amount of data in real time. This requires hardware and software that is capable of processing these data with the least amount of downtime. The processing power and data storage size required increases as the number of mobile users are involved and also as the area of the mobile environment enlarges.

Future challenges in this area are the development of hardware that are capable to process faster, as well as software that can filter the irrelevant data from mobile users and also summarizes the data from mobile users in order to provide a shorter input to the processing system. Other strategies for such issues include the simplification of technique or performance enhancement of techniques in order to reduce the resources required to process such data.

For example, consider the sources of gathered data from many different places. These data can easily contain error, or be duplicated

and misrecorded. To improve performance and accuracy, if there is a probability of error associated with the data, the system could be designed to drop the data, not taking it into consideration and thus improving data accuracy and performance at the same time. Filtering techniques could also be provided to intelligently identify and filter through dropping data or summarizing data. Consider there were duplicate data about Adam, but they are slightly different. Through cross-checking with another data source, the system is able to properly identify whether the data are reliable or whether there is more than one Adam.

CONCLUSION

The future of data mining can be enhanced by capitalizing the increasing use of mobile phones. This gives another mobile dimension of data analysis which produces knowledge that is of higher accuracy. The future use of mobile user data mining technology is described in three different models—the marketing industry model, the banking industry model, and the retail industry model.

The marketing industry and banking industry models are highly centralized. The retail industry model demonstrates the use of mobile user data mining in a distributed manner. Overall, the main benefits provided by mobile user data mining are:

- a. reduced searching cost of consumer data;
- b. ease of tracking, updating, and sharing consumer knowledge; and
- c. the ability to generate knowledge out of a wide amount of data sources with the use of technology that enhances the speed of the process.

The main challenges for mobile user data mining in the future are privacy issues and

knowledge security issues. The research into mobile user data mining continues to provide insights of solutions in different areas, such as security, privacy, and implementation architecture.

REFERENCES

- Agrawal, R., & Srikant, R. (1994). *Fast algorithms for mining association rules*. Santiago: Morgan Kaufmann.
- Agrawal, R., & Srikant, R. (1995). *Mining sequential patterns*. Taiwan: IEEE Computer Society Press.
- Christophides, V., Karvounarakis, G., & Plexousakis, D. (2003). Optimizing taxonomic Semantic Web queries using labelling schemes. *Journal of Web Semantics*, 1(2), 207-228.
- Dourish, P. (2004). What we talk about when we talk about context. *ACM Journal of Pervasive and Ubiquitous Computing*, 8(1), 19-30.
- Eirinaki, M., & Vazirgiannis, M. (2003). Web mining for Web personalization. *ACM Transactions on Internet Technology*, 3(1), 1-27.
- Goh, J., & Tanar, D. (2004a, December 13-15). An efficient mobile data mining model. *Proceedings of 2nd International Symposium on Parallel and Distributed Processing and Applications*, Hong Kong (LNCS, pp. 54-58). Berlin: Springer-Verlag.
- Goh, J., & Tanar, D. (2004b, December 20-25). Mining frequency pattern from mobile users. *Proceedings of the Knowledge-Based Intelligent Information & Engineering and Systems* (LNCS, Part III, pp. 795-801). Berlin: Springer-Verlag.
- Goh, J., & Tanar, D. (2004c, August 25-27). Mining physical parallel pattern from mobile users. *Proceedings of the International*

- Conference on Embedded and Ubiquitous Computing*(LNCS,pp.324-332). Berlin: Springer-Verlag.
- Goh, J., & Tanian, D. (2004d, August 25-27). Mobile data mining by location dependencies. *Proceedings of the 5th International Conference on Intelligent Data Engineering and Automated Learning*, Exeter (LNCS, pp. 225-231). Berlin: Springer-Verlag.
- Goh, J., & Tanian, D. (2005, January-March). Mining parallel pattern from mobile users. *Proceedings of the International Journal of Business Data Communications and Networking* (Vol. 1, No. 1, pp. 50-76). Hershey, PA: Idea Group Publishing.
- Han, J., Dong, G., & Yin, Y. (1999). *Efficient mining of partial periodic patterns in time series database*. Sydney, Australia: IEEE Computer Society.
- Han, J., Gong, W., & Yin, Y. (1998). *Mining segment-wise periodic patterns in time related databases*. Menlo Park: AAAI Press.
- Han, J., Pei, J., & Yin, Y. (2000). *Mining frequent patterns without candidate generation*. New York: ACM Press.
- Kastaniotis, G., Zacharis, N., Panayiotopoulos, T., & Douligeris, C. (2004, May 5-8). Intelligent Web prefetching based upon user profiles—the WebNaut case. *Proceedings of the Conference on Lecture Notes in Artificial Intelligence*, Samos (pp. 54-62).
- Kim, B. J., Kim, I. K., & Kim, K. B. (2004, May 26-28). Feature extraction and classification system for nonlinear and online data. *Proceedings of the 8th Pacific-Asia Conference on Knowledge Discovery and Data Mining 2004*, Sydney, Australia (pp. 171-180).
- Koperski, K., & Han, J. (1995). *Discovery of spatial association rules in geographical information databases*. London: Springer-Verlag.
- Li, J., Tang, B., & Cercone, N. (2004, May 26-28). Applying association rules for interesting recommendations using rule templates. *Proceedings of the 8th Pacific-Asia Conference on Knowledge Discovery and Data Mining 2004*, Sydney, Australia (pp. 166-170).
- Lim, E.-P., Wang, Y., Ong, K.-L., Hwang, S. Y. (2003). In search of knowledge about mobile users. *ERCIM News*, 1(54), 10.
- Miyahara, T., Suzuki, Y., Shoudai, T., Uchida, T., Takahashi, K., & Ueda, H. (2004, May 26-28). Discovery of maximally frequent tag tree patterns with contractible variables from semi-structured documents. *Proceedings of the 8th Pacific-Asia Conference on Knowledge Discovery and Data Mining 2004*, Sydney, Australia (pp. 133-144).
- Oliveira, S. R. M., Zaiane, O. R., & Saygin, Y. (2004, May 26-28). Secure association rule sharing. *Proceedings of the 8th Pacific-Asia Conference on Knowledge Discovery and Data Mining 2004*, Sydney, Australia (pp. 74-85).
- Thiruvady, D. R., & Webb, G. I. (2004, May 26-28). Mining negative rules using GRD. *Proceedings of the 8th Pacific-Asia Conference on Knowledge Discovery and Data Mining 2004*, Sydney, Australia (pp. 161-165).
- Wang, Y., Lim, E.-P., & Hwang, S.-Y. (2003, May 26-28). On mining group patterns of mobile users. *Proceedings of the Conference on Database and Expert System Applications*, Sydney, Australia (pp. 287-296).
- Yip, A. M., Wu, E. H., Ng, M. K., & Chan, T.F. (2004). An efficient algorithm for dense regions discovery from large-scale data streams. *Proceedings of the 8th Pacific-Asia Conference on Knowledge Discovery and Data Mining 2004*, Sydney, Australia (pp. 116-120).

Chapter XVII

Dynamic Matching of Supply and Demand in an M-Commerce Services Marketplace: Using Intelligent Agents and Semantic Technology

Thomas Leary

University of North Carolina at Greensboro, USA

A. F. Salam

University of North Carolina at Greensboro, USA

Rahul Singh

University of North Carolina at Greensboro, USA

ABSTRACT

Mobile users desire customized bundles of services that need to be dynamically created from the service providers. However, services are unique and since unused services do not generate revenues they present a lost 'economic rent' for organizations that are not part of the network of service providers and, as a result, not a part of the customised bundle of services. The dynamic discovery of a bundle of individual services from such a network that meets the unique needs and constraints of the mobile user requires intelligent agent technology. Such agent technology would match personal needs of the user with the available services in a cost-efficient manner. This chapter provides a mechanism to create dynamic service bundles from ad-hoc user requirements using intelligent agents. The authors apply this technique to a mobile commerce environment and illustrate the composition of user-specific service "bundles" by intelligent agents that represent the interests of the m-commerce user. Such agent-based architectures provide users customized solution 'bundles' that reduce their cognitive burden, while improving the utilization of resources for organization that are part of the service provider network.

INTRODUCTION

Forecasts show that mobile Internet users were to grow to 729 million in 2005 (Juptner, 2002), which translates into worldwide mobile commerce (m-commerce) revenue growing from \$400,000 in 2000 to over \$22 billion in 2005 (Macklin, 2001). Currently, mobile Internet users make up 16% of the total Internet users worldwide, with the number expected to grow to near 57% of total Internet users by 2007 (Magura, 2003). With the growing use of mobile devices by consumers to purchase products and services, businesses have an opportunity to meet the unique needs of mobile consumers.

M-commerce is described as the purchase of goods and services via the Internet—much like e-commerce—using mobile devices such as mobile phones and personal digital assistants (PDAs) as the interface. The widespread adoption of mobile phones in the United States, with 140 million phone subscribers in 2002 (Kurzweil, 2004), presents a significant consumer base. Consumers using portable wireless devices such as mobile phones or PDAs for m-commerce activities are often constrained by location and time dependencies (Magura, 2003). A major hindrance to mobile consumers is the low processing power on portable devices to compute and display large amounts of information related to services of interest.

Services

Unlike products, services are intangible goods, but are solutions to consumers' needs and requirements. Services have distinct characteristics, such as location and time dependencies that must be considered in matching consumers and service providers. Services have a temporal aspect where the value is realized, where product's value can be consumed at many times. Products can be shipped from across the world, but services are bound to a

general location. Consumers will only consider services provided in their general area due to travel time and effort required to consume services offered in disparate locations.

Service providers have unique needs, since unused services do not generate revenues and constitute lost 'economic rent' for the provider organizations. For example, in airline seat reservations, the airplane still uses the same amount of fuel whether or not the seat contains a passenger, but the empty seat brings no revenue to defray fixed costs. Another example is a hotel that has rooms available, but still has to pay staff the same amount of money whether the hotel is completely booked or only half-full.

Consumers in need of services, and businesses offering those services, require a way to efficiently communicate their specific needs and capabilities required to fulfill each requested service. In addition, consumers often require a set of services that go beyond the individual capabilities of a single organization (such as food, a taxi, and a hotel room). Such a collection of services, referred to herein as *service bundles* that are sets of individual capabilities of service providers that represents a closer match to the consumers, need then the current process.

The concept of creating a repository for services that facilitates easier discovery is not new. Existing research shows the ability to create ontology to assist in the discovery of services (Hopmans, Klundert, Braspenning, & Kruijsen, n.d.). However, this research does not perform the creation of bundles that truly satisfy mobile users' demands. The AgentCities project is an effort to explore the use of agents to discover and communicate between one another (Agentcities.RTD, 2005). The AgentCities initiative has sparked interest in the area of using agents to support m-commerce activities (Sadeh, Chan, & Van, 2002). However, this research has been focused on suggesting one particular service based on the

location of the consumer. Other research builds upon the AgentCities project; however, they suffer from a lack of much structure or many options (Dale & Ceccaroni, 2002).

In this chapter we present an architecture that supports the dynamic composition of both the consumers' requirements as well as the current service providers' offerings. Consumers have the choice of what bundle best meets their needs, while constantly updating their customer profile to better match with their needs.

BACKGROUND

M-Commerce

M-commerce users have unique goals that arise from time and location pressures (Venkatesh, Ramesh, & Massey, 2003; Magura, 2003; Kakihara & Serensen, 2002). The time pressure is because mobile users will not and perhaps cannot utilize their wireless devices for search-intensive activities due to the technical limitations of bandwidth and lack of processing power. Instead, m-commerce consumers need Web sites that provide solutions for time-critical activities and take into account the location dependence of services. In addition, the time pressure is in part due to the location pressure limiting the availability of choices. Consumers find themselves searching for services using search engines or portals. Service providers must find ways to draw consumers toward their offerings through marketing efforts or pay-per-visit agreements with portals. However, the user must bear the burden of finding a service to meet his/her needs.

Consumers often require a set of services that go beyond the individual capabilities of a single service providing organization, such as food, a taxi, and a hotel room. The dependence of one service on another presents a challenge for users to find services that meet their speci-

fications, as well as the location and time commitments with respect to the other services. These service bundles represent a closer match to consumers' need and are sets of individual capabilities of service providers.

Intelligent Agents

Intelligent software agents can facilitate the discovery and integration of e-business processes, and manage information and knowledge flows for participating firms. An intelligent agent is "a computer system situated in some environment and that is capable of flexible autonomous action in this environment in order to meet its design objectives" (Jennings & Wooldridge, 1998). Intelligent agents are action-oriented abstractions in electronic systems, entrusted to carry out various generic and specific goal-oriented actions on behalf of users. The agent paradigm can support a range of decision-making activities including information retrieval, generation of alternatives, preference order ranking of options and alternatives, and supporting analysis of the alternative-goal relationships. In this respect, intelligent agents have come a long way from being digital scourers and static filters of information to active partners in information processing tasks. Such a shift has significant design implications on the abstractions used to model information systems, objects, or agents, and on the architecture of information resources that are available to entities involved in the electronic system.

Klusch (2001) points out that the specific autonomous behavior expected of intelligent agents depends on the concrete application domain and the expected role and impact of intelligent agents on the potential solution for a particular problem that the agents are designed to provide cognitive support. Muller (1997) identifies three minimal criteria of the applica-

tion domain to apply agent technology. According to Muller, the application domain should:

1. exhibit natural distributivity, e.g., autonomous entities, geographical distribution, and distributed data;
2. require flexible interaction, e.g., there is no a priori assignment of tasks to actors and there are no fixed processes; and
3. be embedded in a dynamic environment, for example, our physical world, artificial worlds like the Internet, and the world of finance.

E-marketplace systems are distributive, flexible, and dynamic environments well suited for the application of agents (Kang & Han, 2002). Multi-agent systems can be used to enhance information flows among e-marketplace participants and work with human agents to organize, store, retrieve, search, and match information and knowledge for e-commerce and m-commerce users.

Semantic Web

The Semantic Web is an extension of the current Web in which information is given “well-defined meaning” to allow machines to “process and understand” the information presented to them (Berners-Lee, Hendler, & Lassila, 2001). This framework provides the basis of our architecture to bundle services dynamically. Bundling of services requires structure to categorize and describe the available services offered in the marketplace. Categorization of services is possible by describing services using an agreed upon ontology. Requests for services can be understood by a reasoning engine and compared to description of services available. While all services must contain the same general information including time restrictions and price, further specialization can be utilized in individual vertical markets

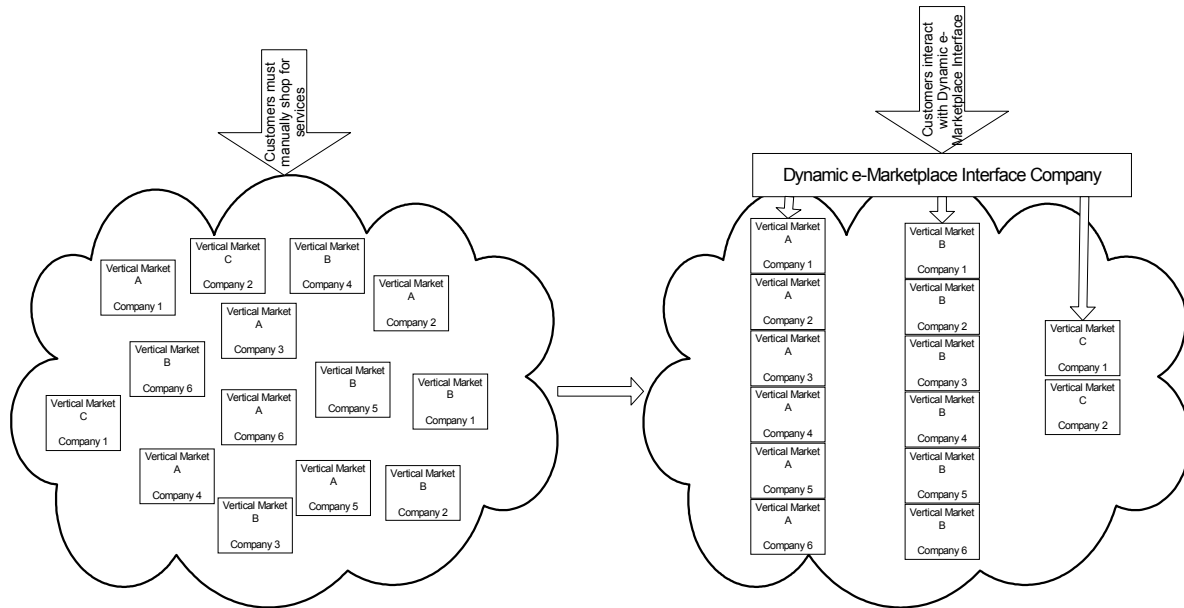
to provide more detailed information depending on the particular service offering. All vertical markets can specify information that is important in comparing similar services that meet consumer demand, but are not necessary in order to compose services into bundles.

Discovery and comparison of different services by intelligent agents requires a common vocabulary to exist to describe the individual services. The “Semantic Web” has a computer-interpretable language to describe the capabilities and content of a particular service (McIlraith & Martin, 2003). This common language allows heterogeneous data from various sources to be integrated to “perform sophisticated and detailed analysis” (Michalowski et al., 2004). Description Logics is the foundation describing and providing a hierarchical structure to concepts. Recent academic papers have used DAML+OIL as their language (McIlraith & Martin, 2003). A recent standard adopted by the W3C is OWL-S, which “supplies a core set of markup language constructs for describing the properties and capabilities of Web services in unambiguous, computer-interpretable form” (W3C, 2005). This allows discovery, execution, composition, and interoperation of the services. Both DAML+OIL and OWL use Resource Description Framework as their syntax. Therefore, according to Sycara, Paolucci, Ankolekar, and Srinivasan (2005), OWL can be substituted for DAML+OIL whenever referenced in literature with little or no modification.

PROPOSED INTELLIGENT SERVICES MARKETPLACE FOR M-COMMERCE

In this work, we present a potential solution to the current inefficiencies inherent in the service provider marketplace as described in the previous section. Consumers and the service providers need infrastructure that provides ac-

Figure 1. Transition of current m-commerce to a more dynamic structure



cess to each other to meet their needs. M-commerce consumers need an intelligent marketplace to reduce their cognitive and technological load in the buying process (Maamar, 2003; Kotz & Gray, 1999). Service providers need to have a dynamic marketplace, where available services can be discovered and presented to as many customers as possible.

E-marketplace models can be made more efficient by utilizing intelligent agents to work on behalf of e-marketplace users to serve their needs in an information-intensive environment (Papazoglou, 2001). For this research, an e-marketplace provides the necessary information-sharing infrastructure to allow service providers to publish their service offerings inside an organized repository. This creates an environment where agents can process the information in the repository to discover individual or groups of services customized to meet the needs of the m-commerce consumers. In addition, customized service bundles created by

intelligent user-agents, with knowledge of user preferences, reduce the information overload of consumers by eliminating the need for exhaustive searching for individual services. Businesses gain access to these mobile users by publishing intelligent agents into an e-marketplace with knowledge of their available services, providing an avenue for better utilization of service resources for increased revenue generation.

The dynamic discovery of aggregated service bundles from the individual offerings of a network of service providers to meet the unique needs and constraints of a mobile user requires intelligent agent technology to match the user needs with the available services. This chapter presents a mechanism to create dynamic service bundles from ad-hoc user requirements using intelligent agents. We illustrate the composition of user-specific service bundles by intelligent agents that represent the interests of the m-commerce user in the marketplace. Such

agent-based architectures provide users customized solution bundles. This reduces their cognitive burden by sparing them from manual searches of available services in a local area. Companies that are members of this network enjoy improved utilization of resources and increase potential revenues by selling otherwise *perishable* services that would provide the organization no income and incur fixed costs.

REQUIREMENTS OF AN INTELLIGENT E-MARKETPLACE

Mobile commerce must provide the user with information about locally relevant services by capturing the location of the user. This location specificity is important in providing consumers useful information instead of the information overload that plagues regular e-commerce consumers (Venkatesh et al., 2003). In addition, the mobile environment does not have the computing resources or the bandwidth to send large amounts of data to accommodate large numbers of choices for the user to consider. Therefore, determining what the user would value based on the location, previous actions, and personal preferences requires logic to limit overload and presents the user with the offerings closest matching their criteria and preferences.

The consumer situation can be observed, analyzed, and used to provide a mobile user with information that meets the current situation in which the user finds himself/herself (Figge, 2002). This process uses three main situational components: *identity*, *position*, and *time*. The identity situation captures and allows processing of the user's personal profile to reflect his/her personal characteristics. For example, the personal profile can contain the information that this customer prefers three-star hotels, as well as restaurants that serve

entrées that cost around \$20. The location profile for the user provides the current location to be used to understand more about the situation, such as whether the user is in the office or in a recreational venue. The time situation information provides information specific to predicted consumer needs based on that time. For example, the time profile can remember that this user prefers to eat around 7:00 p.m. and usually goes to bed around 11:00 p.m. Based on this, the e-marketplace can provide a specific set of situation-dependent services that meet the personalized needs of that particular user. Mobile users benefit from an arrangement of services to meet their time and space requirements, while considering their own personal preferences.

This sequential arrangement or composition of service bundles means that the output of one service becomes the input for another service (Berger et al., 2003). This is a key concept in that a system that features service composition can be extremely flexible to accomplish the demands put on that system.

We apply the Service-Oriented Process (SOP) model to the e-marketplace to compose a set of services to meet the unique needs of the mobile users. We utilize Georgakopoulos, Schuster, Cichocki, and Baker's (2002) list of properties of an SOP model: (1) capture products/services from vendors from many different vertical markets, and select and integrate the services into different product groupings or service compositions; (2) capture e-services—model the functional, contractual, and conversational relationships between the service providers in the marketplace; and (3) allow for the composition of services for mobile commerce users by automatic coordination and organization of grouping of products/services desired by the m-commerce user (Georgakopoulos et al., 2002). Table 1 shows the requirements of an agent-enabled e-marketplace for m-commerce users.

Dynamic Matching of Supply and Demand in an M-Commerce Services Marketplace

Table 1. Requirements of agent-enabled m-commerce e-marketplaces

Wireless Coverage Availability	While urban areas are generally well covered, rural areas still have uncovered digital areas needed to utilize the m-commerce system.
Multiple Service Providers	A network of service providers to meet consumers' needs. Similar to the "network" effect, where the more businesses participating in the marketplace, the more valuable it becomes for providers, consumers, and the network.
Automatic Discovery, Matching, and Composition of Services	Services need to be automatically discovered, matched to specific requests, and aggregated into bundles of services through "Matchmaker" capabilities.
Ontology	For the marketplace to aggregate and search through different choices, standard accepted vocabulary for describing services must be used (Pan, Cranefield, & Carter, 2003; Williams, Padmanabhan, & Burke, 2003).
Preference Storage and Adaptation	Users must be able to set their preferences for services and to customize these preferences over time to better meet their changing needs.
Location	Location must be able to be determined with some accuracy. Technology such as E-911 and GPS is being built into most phones and can be added to many PDAs to meet this requirement.
Trust/Security	Digital signatures, encryption mechanisms, and authorization functionality can offer consumers confidence in wireless communication security, which in turn boosts trust in m-commerce (Siau & Shen, 2003).
Proof of Purchase	The system must provide consumers with a proof of purchase that will prove to the business that the transaction has occurred. Methods of payment can include the wireless service provider bill, direct to credit card, or straight out of bank account (Mallat, Rossi, & Tuunainen, 2004; Herzberg, 2003).
Decision Time Allotment	Businesses must allow the consumer time to check the options available and make a decision. However, this creates problems with businesses reserving products/services for wireless users when they could have sold them to other customers.

Figure 2. An agent-enabled architecture to model the information flows required for service bundle composition

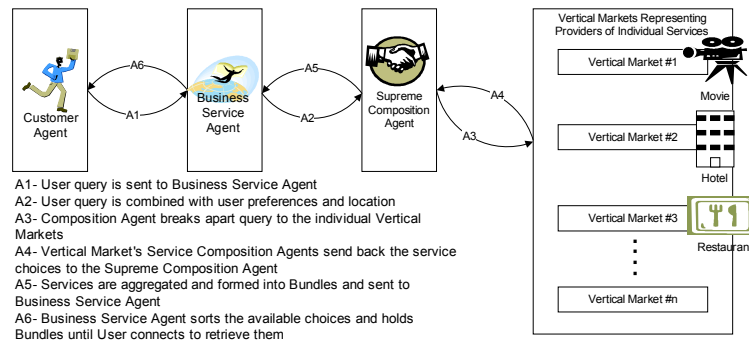
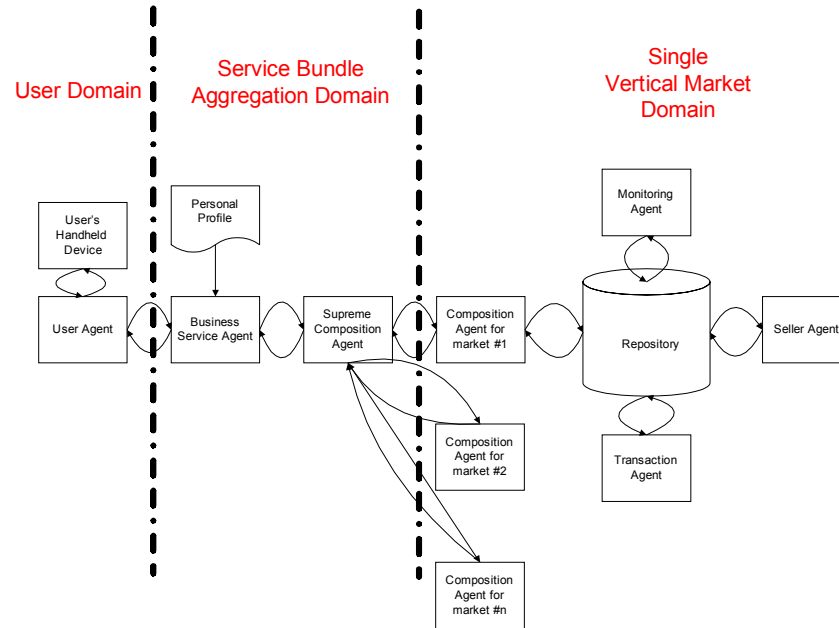


Figure 3. A specific view of an agent-enabled architecture from which to model the information flows required for service bundle composition



PROPOSED SERVICE ARCHITECTURE

By ascertaining the time and location demands of the user, the architecture provides easy access to time-saving, location-specific services. The user can consider, view alternatives, and securely purchase in a manner that meets the unique characteristics of the mobile technology such as global positioning systems (GPSs) with which many cell phones and PDAs can be (or already are) equipped. After receiving the location of the user, agents are able to utilize past buying behavior and certain customer preferences in order to provide bundles of services that represent a smaller set of applicable choices. This elimination of services irrelevant to either the search criteria or buyer preferences cuts buyer search costs, while possibly exposing customers to new vendors that can better meet their needs.

Figure 2 illustrates the conceptual model of how the agent architecture can take in the location from the user and disseminate the information throughout the e-marketplace. The user submits the query for a group of services to the business service agent. This agent combines the query with personal preferences that are located in the individual's personal profile. The individual's preferences about the particular services requested are combined and sent onto the Supreme Composition Agent. This agent then breaks the query with preferences into different parts to their respective vertical markets. In Figure 2, we see that the user requests services from the movie, hotel, and restaurant vertical markets. In step A4, the services available are sent back to the supreme composition agent from each vertical market, where the services are grouped into bundles and pushed back to the user.

Figure 4. Illustration of composition in vertical markets

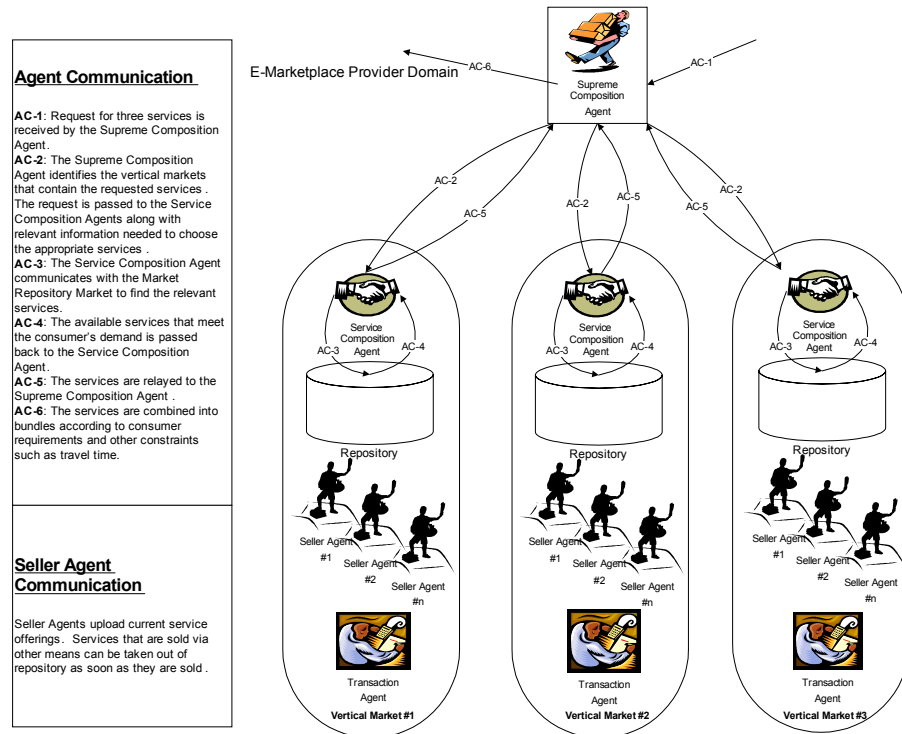
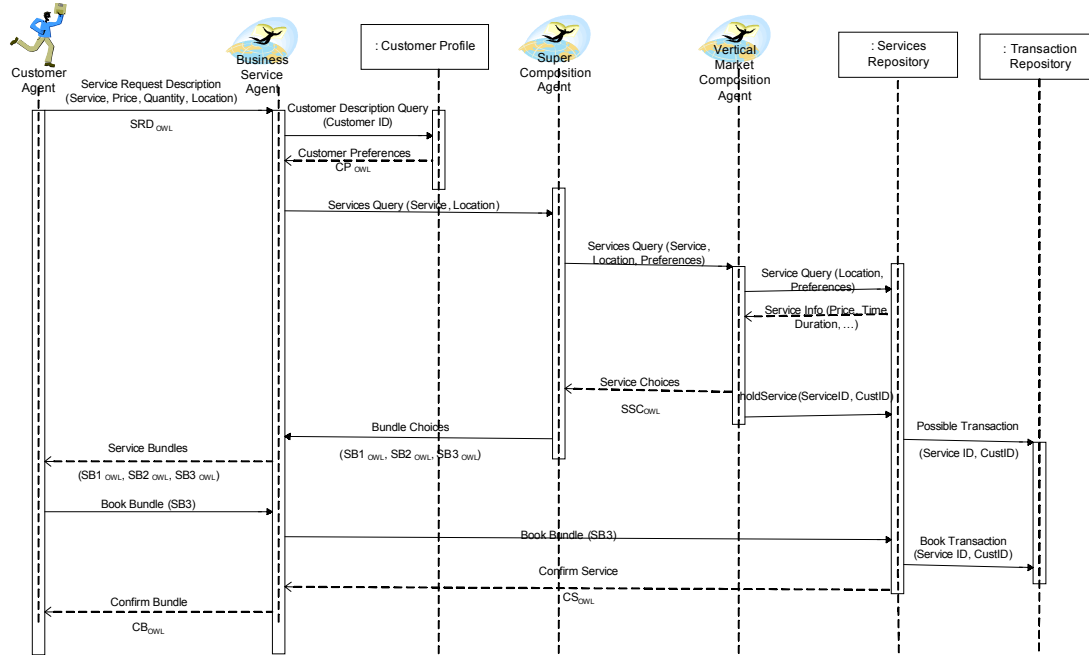


Figure 3 shows the specific illustration of one particular vertical market, such as the hotel vertical market from Figure 2. The composition agents in the bundle service aggregation domain interact with one discovery agent per vertical market. Therefore, the supreme composition agent would interact with a composition agent from each vertical market. In this multiple vertical market scenario, the supreme composition agent must split the query to provide the necessary information to the respective vertical markets in order to find relevant services. In addition, when the composition agents return the available services that meet the demands of the user, the supreme composition agent aggregates services into combinations, taking into account the time factors of the

individual services. These bundles are passed back to the consumer for acceptance.

The Service Composition Provider owns all agents and acts as an intermediary between the service providers and the consumer. This is shown as the Service Bundle Aggregation Domain in Figure 2. In the m-commerce example, the Service Composition Provider is the owner of the wireless network, such as Verizon Wireless, Cingular, and so forth. The User Agent is responsible for the interaction with the consumer, whereas the service providers interact with the Seller Agent who publishes the relevant information to the respective vertical market. The rest of the infrastructure is transparent to the consumer and the service providers.

Figure 5. Sequence diagram of service bundle composition



The Supreme Composition Agent breaks down the request for services for use of the respective Service Composition Agents. The Supreme Composition Agent receives the request for three services in Figure 4. Next, the Supreme Composition Agent determines the vertical market serving the needs of that particular service and contacts the Service Composition Agents representing those vertical markets. In each of the vertical markets, the Service Composition Agent scours the repository for services currently being offered by the many sellers operating inside that vertical market. If services that meet the buyer's requirement are found, these services are passed back to the Supreme Composition Agent. The Supreme Composition Agent takes these possible services from the three vertical markets and creates possible bundles.

Service Publishing and Inference Mechanism

We have described how services located in a vertical market businesses need to “publish” their available services in the e-marketplace. Universal Description Discovery and Integration (UDDI) allows businesses to show their availability in the marketplace and provide a description of the services they provide. UDDI permits an unlimited set of properties to the description of the Web Service. In the proposed architecture, service properties are specified using specific ontology for a vertical market. Specifying the standard definitions of service properties allows for syntactic interoperability; however, the semantics are not specified. Semantic Web services can be achieved in utilizing the OWL-S virtual machine (Sycara, 2005).

Our proposed architecture shown in Figure 3 uses the Web-service invocation, the OWL-S processor, and the OWL inference engine. The Web-service invocation handles the communications between the UDDI registry, which then passes it to the OWL-S processor. The processor translates it into OWL, and sends this to the OWL inference engine that interprets the message according to a specified ontology. The inference engine makes a decision based on the information available from the mobile consumer's request and the available services located in the vertical market. In Sycara et al.'s (2003) model, this inference engine, designated the "Matchmaker," selects the service offerings that are relevant for a request. In Noia's work (Noia, Sciascio, Donini, & Mongiello, 2003), this inference mechanism is called the "MatchMaker service," which can compute matches' classification and ranking. Using an inference mechanism that performs both composition and ranking of possible bundles allows the user to gain a greater understanding and context on the offerings presented.

Service Composition

We present a sequence diagram to show the interaction between different agents to allow for the collection of a query from a mobile consumer, working with their personal profile, and then finding the relevant services that meet their needs in order to create a series of bundles.

The customer begins the process for declaring his/her request for a group of services. In this request, the services requested, price range, quantity, and location desired are passed to the Business Service Agent. The customer's individual preferences for these types of services are found in the customer profile and added to the query. The Business Service Agent sends the complete query to the Super Composition Agent (SCA), which breaks apart the respec-

tive pieces of the query in order to send to the appropriate Vertical Market Composition Agent. Each Vertical Market Composition Agent queries the database to find services in that marketplace that meet the location and customer preferences. The matches are passed back to the Vertical Market Composition Agent, which takes the "best" possible choices back to the Super Composition Agent. This SCA gathers responses from each of the vertical markets queried. Now, the SCA can decide which services can be bundled together to best meet the criteria of the customer. The services present in the bundle are reserved for a period of time to allow for acceptance/rejection. The bundle listing is passed back to the customer, which chooses for a particular bundle or rejects all. If the consumer accepts a bundle, the constituent services are booked and the confirmation of this is passed back to the consumer.

CONCLUSION AND FUTURE DIRECTIONS

A variety of payment options exist following the bundle consumption. In the first scenario, service providers bill the consumer directly as another line item on their wireless service bill (Herzberg, 2003). Details to the services consumed are available for the consumer to double-check. The second scenario involves the integration of a credit card company, where the customer enters the credit card information as soon as a bundle is booked (Mallat et al., 2004). Regardless of the method used, a common theme emerges in that users must feel secure with the security precautions of these payment options, as well with safeguards in place ensuring payment to service providers.

We have illustrated the need for m-commerce to represent the current offerings of service providers and the preferences of a

mobile consumer in order to provide personalized bundles of services. Our architecture builds on previous research that provides syntactic interoperability and the semantic interoperability (Sycara et al., 2003). Syntactic interoperability permits agents to understand the structure of the communication being given, while semantic interoperability allows agents to understand the context and meaning of the message, primarily by Sycara et al. (2003) and McIlraith, Son, and Zeng (2001).

In Berger et al.'s (2003) work, m-commerce is shown to become the next step in electronic commerce due to the demands of the mobile society. People are constantly on the move, utilizing cell phones and PDAs to connect wirelessly to Internet resources. Yet, people do not have the convenience of effectively discovering relevant services efficiently. The current infrastructure does not meet the demands of mobile consumers and the mobile hardware due to its forcing the consumer to wade through Web searches, localized Web sites (e.g., CitySearch), or waiting to access a desktop computer. The proposed architecture provides consumers with easy access to services without information overload from anywhere, even with intermittent wireless connection. Consumers specify personal preferences in order to narrow the groupings of services. Businesses, on the other hand, gain access to more consumers to attempt to sell the available perishable services.

While this model hinges on several important factors, the "network" effect needed a critical mass of businesses in order to provide consumers with an adequate selection of services in different areas. This model would best be implemented in major markets first, then spreading to smaller markets in order to capture a large amount of service providers during the initial phase. Future research should be directed towards the inference mechanism for

determining relevant services, security measures of the system, and the user interaction (user interface and the communication between the wireless device and the m-commerce system). Once there is an inference mechanism for the intelligent agents in the architecture, work can be started on a prototype system to show the service groupings that the system would produce to the consumer given a location and preferences.

REFERENCES

- Agentcities.RTD. (n.d.). Retrieved from <http://www.agentcities.org/EURTD/>
- Berger, M., Bouzid, M., Buckland, M., Lee, H., Lhuillier, N., Olpp, D., Picault, J., & Shepherdson, J. (2003). An approach to agent-based service composition and its application to mobile business processes. *IEEE Transactions on Mobile Computing*, 2, 197-206.
- Berners-Lee, T., Hendler, J., & Lassila, O. (2001, May). The Semantic Web. *Scientific American*, 34-43.
- Dale, J., & Ceccaroni, L. (2002). Pizza and a movie: A case study in advanced Web services. *Proceedings of AAMAS'02*.
- Figge, S. (2002). Situation-dependent services—A challenge for mobile network operators. *Journal of Business Research*, 58.
- Georgakopoulos, D., Schuster, H., Cichocki, A., & Baker, D. (2002). Process-based e-service composition for modeling and automating zero latency supply chains. *Information Systems Frontiers*, 4, 33-54.
- Herzberg, A. (2003). Payments and banking with mobile personal devices. *Communications of the ACM*, 46, 53-58.

- Hopmans, G., Klundert, J., Braspenning, P., & Kruijsen, P.-P. (n.d.). *Maastricht-restaurant selection and reservation-services*.
- Jennings, N.R., & Wooldridge, M. (1998). *Agent technology: Foundations, applications, and markets*. London: Springer.
- Juptner, O. (2004). Retrieved February 20, 2004, from <http://www.e-gateway.net/infoarea/news/news.cfm?nid=2131>
- Kakihara, M., & Serensen, C. (2002). Mobility: An extended perspective. *Proceedings of the 35th HICSS Conference*.
- Kang, N., & Han, S. (2002). Agent-based e-marketplace system for more fair and efficient transaction. *Decision Support Systems*, 34, 157-165.
- Klusck, M. (2001). Information agent technology for the Internet: A survey. *Data and Knowledge Engineering*, 36, 337-372.
- Kotz, D., & Gray, R.S. (1999). Mobile agents and future of the Internet. *ACM Operating Systems Review*, 33, 7-13.
- Kurzweil, R. (2004, February). *Harvard Business Review*, 37.
- Maamar, Z. (2003). Commerce, e-commerce, and m-commerce: What comes next? *Communications of the ACM*, 46, 251-257.
- Macklin, B. (2004). Retrieved February 20, 2004, from <http://www.entrepreneur.com/article/0,4621,290409,00.html>
- Magura, B. (2003). What hooks m-commerce customers? *MIT Sloan Management Review*, (Spring).
- Mallat, N., Rossi, M., & Tuunainen, V. K. (2004). Mobile banking services. *Communications of the ACM*, 47, 42-46.
- McIlraith, S. A., & Martin, D. L. (2003, January/February). Bringing semantics to Web services. *IEEE Intelligent Systems*, 90-93.
- McIlraith, S. A., Son, T. C., & Zeng, H. (2001, March/April). Semantic Web services. *IEEE Intelligent Systems*, 46-53.
- Michalowski, M., Ambite, J. L., Thakkar, S., Tuchinda, R., Knoblock, C. A., & Minton, S. (2004, May/June). Retrieving and semantically integrating heterogeneous data from the Web. *IEEE Intelligent Systems*, 2-79.
- Muller, H. J. (1997). Toward agent systems engineering. *Data and Knowledge Engineering*, 23, 217-245.
- Noia, T. D., Sciascio, E. D., Donini, F. M., & Mongiello, M. (2003). A system for principled matchmaking in an electronic marketplace. *Proceedings of WWW2003*.
- Pan, J., Cranefield, S., & Carter, D. (2003). A lightweight ontology repository. *Proceedings of AAMAS'03*.
- Papazoglou, M. P. (2001). Agent oriented technology in support of e-business: Enabling the development of intelligent business agents for adaptive, reusable software. *Communications of the ACM*, 44, 71-77.
- Sadeh, N. M., Chan, E., & Van, L. (2002). Open agent environment for context-aware m-commerce. *Proceedings of AAMAS'02*.
- Siau, K., & Shen, Z. (2003). Building customer trust in mobile commerce. *Communications of the ACM*, 46, 91-94.
- Sycara, K. (n.d.). Retrieved from <http://www-2.cs.cmu.edu/~softagents/daml.html>
- Sycara, K., Paolucci, M., Ankolekar, A., & Srinivasan, N. (2003). Automated discovery, interaction and composition of Semantic Web services. *Journal of Web Semantics*, 1, 1-28.

Venkatesh, V., Ramesh, V., & Massey, A. P. (2003). Understanding usability in mobile commerce. *Communications of the ACM*, 46, 53-56.

W3C. (2004). Retrieved February 27, 2004, from <http://www.w3.org/2004/OWL/>

Williams, A., Padmanabhan, A., & Burke, M. B. (2003). Local consensus ontologies for B2B-oriented service composition. *Proceedings of AAMAS'03*.

Chapter XVIII

Information Delivery for Mobile Business:

Architecture for Accessing Large Documents through Mobile Devices

Christopher C. Yang

Chinese University of Hong Kong, Hong Kong

Fu Lee Wang

City University of Hong Kong, Hong Kong

ABSTRACT

In this information-centric age, an organization needs to access the most update and accurate information for fast decision making. Mobile access to Internet provides convenient and portable access to a huge information space. However, loading and visualizing large documents on mobile devices is impossible due to their natural shortcomings such as screen size and computing power. In this chapter, we introduce the fractal summarization model, based on fractal theory, for document summarization on mobile devices. This model generates a brief skeleton of summary at the first stage, and the details of the summary on different levels of the document are generated on demands from users. Such interactive summarization reduces the computation load, which is ideal for wireless access. On the other hand, the hierarchical display in fractal summarization is more suitable for navigation of a large document and it is ideal for small area display. The automatic summarization together with the three-tier architecture and the information visualization are potential solutions to the existing problems in information delivery to mobile devices for mobile business.

INTRODUCTION

Access to the Internet through mobile devices is growing significantly in recent years. The wireless application protocol (WAP) and wireless markup language (WML) provide the uni-

versal open standard and markup language. Many information-centric applications have been developed for mobile devices (Buyukkokten, Garcia-Molina, Paepcke, & Winograd, 2000; Buyukkokten, Garcia-Molina, & Paepcke, 2001a, 2001b, 2001c; Yang &

Wang, 2002, 2003b, 2003c). For example, users can now surf the Web, check e-mail, read news, and quote stock prices, using mobile devices. At present, most mobile applications are customer-centered m-services applications. However, mobile computing should not be limited to user-centered applications only. It should be extended to decision support in an m-commerce organization. With a fast-paced economy, organizations need access to large documents or other information sources for fast decision making. As a result, there is an urgent need of a tool for browsing large documents on mobile devices.

Although the development of wireless mobile devices is fast in recent years, there are many shortcomings associated with these devices, such as screen size, bandwidth, and memory capacity. There are two major categories of wireless mobile devices, namely, WAP-enabled mobile phones and wireless personal digital assistants (PDAs). At present, the typical display size of popular WAP-enabled handsets and PDAs is relatively small in comparison with a standard personal computer. The comparatively limited memory capacity of a mobile device also greatly limits the amount of information that can be stored. A large document cannot be entirely downloaded to the mobile device and presented to the user, as the current bandwidth available for WAP is relatively narrow as compared with the broadband Internet connection for PCs.

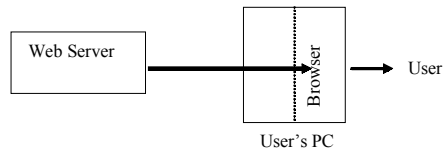
Despite their convenience, mobile devices impose many constraints that do not exist on desktop computers. The low bandwidth and small resolution are major shortcomings of mobile devices. Information overloading is a critical problem; advance-searching techniques solve the problem by filtering most of the irrelevant information. However, the precision of most of the commercial search engines is not high. Users may only find a few relevant docu-

ments out of a large pool of searching results. Given the large screen and high bandwidth for desktop computing, users may still need to browse the search results one by one and identify the relevant information using desktop computers. However, it is impossible to search and visualize the critical information on a small screen with an intolerable slow downloading speed using mobile devices. Automatic summarization summarizes a document for users to preview its major content. Users may determine if the information fits their needs by reading their summary instead of browsing each whole document one by one. The amount of information displayed and downloading time are significantly reduced.

Traditional automatic summarization does not consider the structure of a document, but considers the document as a sequence of sentences. Most of the traditional summarization systems extracted sentences from the source document and concatenated them together as summary. However, it is believed that the document summarization on mobile devices must make use of a "tree view" (Buyukkokten et al., 2001a, 2001b, 2001c) or "hierarchical display" (Mani, 2001). Similar techniques have been applied to Web browsing (Brown & Weihl, 1996): an outline processor organizes the Web page in a tree structure, and the user clicks the link to expand the subsection and view the detail. Hierarchical display is suitable for navigation of a large document, and it is ideal for small area display. Therefore, a new summarization model with hierarchical display is required for summarization on mobile devices.

Summarization on mobile devices in the context of Web pages has been investigated by Buyukkokten et al. (2000, 2001a, 2001b, 2001c). However, a large document exhibits totally different characteristics from Web pages. A Web page usually contains a small number of sentences that are organized into paragraphs,

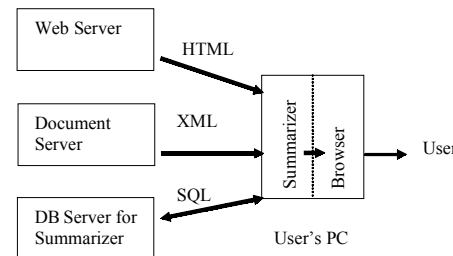
Figure 1. Document browsing on PC



but a large document contains many more sentences that are organized into a more complex hierarchical structure. Also, the summarization on a Web page is mainly based on thematic features only (Buyukkokten et al., 2001a). However, it has been proven that other document features play as important a role as the thematic feature (Edmundson, 1969; Kepiec, Pedersen, & Chen, 1995). Therefore, a more advanced summarization model combined with other document features is required for browsing large documents on mobile devices.

In this chapter, we propose the fractal summarization model based on the statistical data and the structure of documents. Thematic feature, location feature, heading feature, and cue features are adopted. Summarization is generated interactively. Experiments have been conducted, and the results show that the fractal summarization outperforms the traditional summarization. In addition, information visualization techniques are presented to reduce the visual loads. Three-tier architecture, which reduces the computing load of the mobile devices, is also discussed. In addition to large documents, there is a lot of other valuable information available on the Internet. For example, a great amount of financial news is generated everyday. Access to the most updated and accurate financial information is important during decision making. We will demonstrate the financial news delivery on mobile devices as an example of information delivery for mobile business.

Figure 2. Document browsing with summarizer on PC



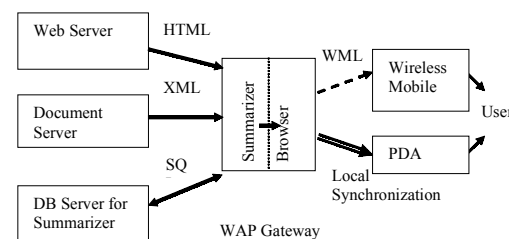
THREE-TIER ARCHITECTURE

Two-tier architecture is typically utilized for Internet access. The user's PC connects to the Internet directly, and the content loaded will be fed to the Web browser and presented to the user as illustrated in Figure 1.

Due to the information-overloading problem, a summarizer is introduced to summarize a document for users to preview before presenting the whole document. As shown in Figure 2, the content will be first fed to the summarizer after loading to the user's PC. The summarizer connects to the database server when necessary and generates a summary to display on the browser.

The two-tier architecture cannot be applied on mobile devices, since the computing power of mobile devices is insufficient to perform

Figure 3. Document browsing with summarizer on WAP



summarization and the network connection of a mobile network does not provide sufficient bandwidth for navigation between the summarizer and other servers.

The three-tier architecture as illustrated in Figure 3 is proposed. A WAP gateway is set up to process the summarization. The WAP gateway connects to the Internet through a broadband network. The wireless mobile devices can conduct interactive navigation with the gateway through a wireless network to retrieve the summary piece by piece. Alternatively, if the PDA is equipped with more memory, the complete summary can be downloaded to the PDA through local synchronization.

AUTOMATIC SUMMARIZATION

Because there are a lot of shortcomings associated with mobile devices, the Traditional Summarization Model cannot be implemented on a mobile device. A novel summarization model based on hierarchical document structure and fractal theory will be presented.

Traditional Summarization

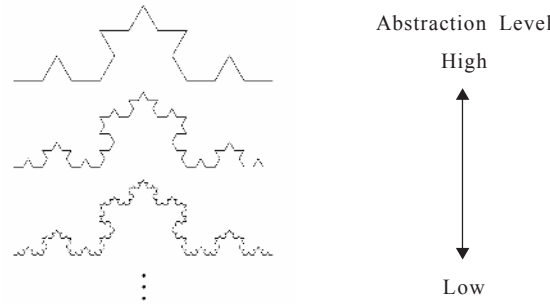
Traditional automatic text summarization is the selection of sentences from the source document based on their significance to the document (Edmundson, 1969; Luhn, 1958). The selection of sentences is conducted based on the salient features of the document. The thematic, location, heading, and cue features are the most widely used summarization features.

- The thematic feature is first identified by Luhn (1958). Edmundson (1969) proposed to assign the thematic weight to keyword based on term frequency and the sentence weight as the sum of thematic weight of constituent keywords. In information retrieval, absolute term frequency by itself

is considered as less useful than term frequency normalized to the document length and term frequency in the collection (Harman, 1992). As a result, the *tfidf* (Term Frequency, Inverse Document Frequency) method is proposed to calculate the thematic weight of keyword (Salton & Buckley, 1988).

- The significance of sentence is indicated by its location (Baxendale, 1958) based on the hypotheses that topic sentences tend to occur at the beginning or end of documents or paragraphs (Edmundson, 1969). Edmundson proposed to assign positive weights to sentences according to their ordinal position in the document—that is, the sentences in the first and last paragraphs and the first and last sentences of the paragraphs. There are several functions proposed to calculate the location weight of sentences. Alternatively, the preference of sentence location can be stored in a list called Optimum Position Policy, and the sentences will be selected based on their order in the list (Lin & Hovy, 1997).
- The heading feature is proposed based on the hypothesis that the author conceives the heading as circumscribing the subject matter of the document. When the author partitions the document into major sections, he summarizes them by choosing appropriate headings (Edmundson, 1969). The formulation of heading weight is very similar to the thematic feature. A heading glossary is a list consisting of all the words in headings and subheadings. Positive weights are assigned to the heading glossary, where the heading words will be assigned a weight relatively prime to the subheading words. The heading weight of a sentence is calculated by the sum of the heading weight of its constituent words.

Figure 4. Koch curve at different abstraction levels



- The cue phrase feature is proposed by Edmundson (1969) based on the hypothesis that the probable relevance of a sentence is affected by the presence of pragmatic words such as “significant,” “impossible,” and “hardly.” A pre-stored cue dictionary is used to identify the cue phrases, which comprises three sub-dictionaries: (i) bonus words, which are positively relevant; (ii) stigma words, which are negatively relevant; and (iii) null words, which are irrelevant. The cue weight of a sentence is calculated by the sum of the cue weight of its constituent words.

Typical summarization systems select a combination of summarization features (Edmundson, 1969; Lin & Hovy, 1997; Luhn, 1958); the total sentence significance score (SSS) is calculated as:

$$SSS = a_1 \times SS_{thematic} + a_2 \times SS_{location} + a_3 \times SS_{heading} + a_4 \times SS_{cue}$$

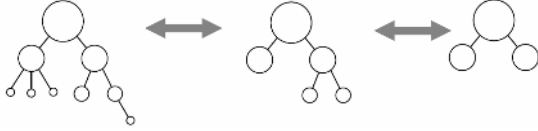
where $SS_{thematic}$, $SS_{location}$, $SS_{heading}$, and SS_{cue} are sentence scores based on thematic feature, location feature, heading feature, and cue phrase feature, respectively, and a_1 , a_2 , a_3 , and a_4 are positive integers to adjust the weighting of four summarization features. The sentences with a sentence significant score higher than a threshold are selected as part of the summary. It has been proved that the weighting of different

summarization features does not have any substantial effect on the average precision (Lam-Adesina & Jones, 2001). In our experiment, the maximum score of each feature is normalized to one, and the sentence significant score is calculated as the sum of scores of all summarization features without weighting.

Fractal Theory and Fractal View for Controlling Information Displayed

Fractals are mathematical objects that have high degree of redundancy (Mandelbrot, 1983). These objects are made of transformed copies of themselves or part of themselves (see Figure 4). Mandelbrot (1983) was the first person who investigated fractal geometry and developed the fractal theory. In his well-known example, the length of the British coastline depends on measurement scale. The larger the scale is, the smaller the value of the length of the coastline is and the higher the abstraction level is. The British coastline includes bays and peninsulas. Bays include sub-bays, and peninsulas include sub-peninsulas. Using fractals to represent these structures, abstraction of the British coastline can be generated with different abstraction degrees. Fractal theory is grounded in geometry and dimension theory. Fractals are independent of scale and appear equally detailed at any level of magnification. Such property is known as self-similarity. Any portion of a self-

Figure 5. Fractal view for logical tree at different abstraction levels

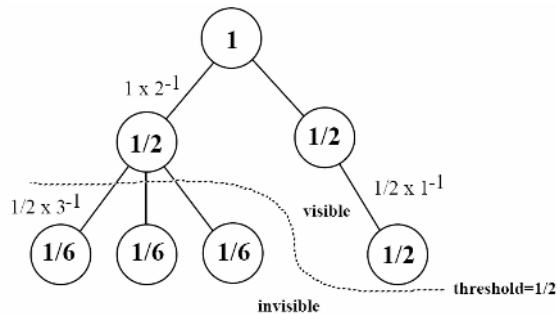


similar fractal curve appears identical to the whole curve. If we shrink or enlarge a fractal pattern, its appearance remains unchanged.

Fractal view is a fractal-based method for controlling information displayed (Koike, 1995). Fractal view provides an approximation mechanism for the observer to adjust the abstraction level and therefore control the amount of information displayed. At a lower abstraction level, more details of the fractal object can be viewed.

A physical tree is one classical example of fractal objects. A tree is made up of many subtrees; each of them is also a tree. By changing the scale, the different levels of abstraction views are obtained (see Figure 5). The idea of fractal tree can be extended to any logical tree. The degree of importance of each node is represented by its fractal value. The fractal value of focus is set to 1. Regarding the focus as a new root, we propagate the fractal value to other nodes with the following expression:

Figure 6. An example of the propagation of fractal values



$$\begin{cases} Fv_{root} &= 1 \\ Fv_{child\ node\ of\ x} &= C \frac{Fv_x}{N_x^{\frac{1}{D}}} \end{cases}$$

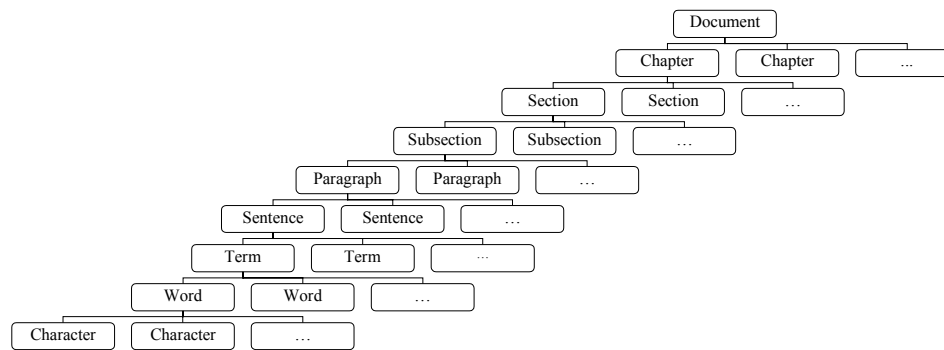
where Fv_x is the fractal value of node x ; C is a constant between 0 and 1 to control rate of decay; N_x is the number of child nodes of node x ; and D is the fractal dimension.

A threshold value is chosen to control the amount of information displayed; the nodes with a fractal value less than the threshold value will be hidden (see Figure 6). By change the threshold value, the user can adjust the amount of information displayed.

Fractal Summarization

Advance summarization techniques take the document structure into consideration to compute the probability of a sentence to be included in the summary. Many studies (Endres-Niggemeyer, Maier, & Sigel, 1995; Glaser & Strauss, 1967) of human abstraction process have shown that the human abstractors extract the topic sentences according to the document structure from the top level to the low level until they have extracted sufficient information. However, most traditional automatic summarization models consider the source document as a sequence of sentences, but ignore the structure of document. Some summarization systems may calculate sentence weight partially based on the document structure, but they still extract sentences in a linear space. In conclusion, none of the Traditional Summarization Models is entirely based on document structure. *Fractal summarization model* is proposed here to generate summary based on document structure. Fractal summarization generates a brief skeleton of summary at the first stage, and the details of the summary on different levels of the document are generated on demand of the users. Such interactive sum-

Figure 7. Prefractal structure of document



marization reduces the computation load in comparing with the generation of the entire summary in one batch by the traditional automatic summarization, which is ideal for m-commerce.

Fractal summarization is developed based on the fractal theory. In our fractal summarization, the important information is captured from the source text by exploring the hierarchical structure and salient features of the document. A condensed version of the document that is informatively close to the original is produced iteratively using the contractive transformation in the fractal theory. Similar to the fractal geometry applying on the British coastline where the coastline includes bays, peninsulas, sub-bays, and sub-peninsulas, a large document has a hierarchical structure with several levels, chapters, sections, subsections, paragraphs, sentences, and terms. A document considered as *prefractal* that is a fractal structure in the early stage with finite recursion only (Feder, 1988).

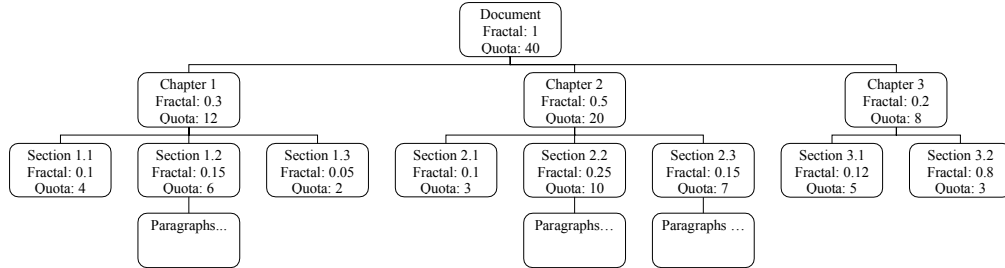
A document can be represented by a hierarchical structure as shown in Figure 7. A document consists of chapters. A chapter consists of sections. A section may consist of subsections. A section or subsection consists of paragraphs. A paragraph consists of sentences. A

sentence consists of terms. A term consists of words. A word consists of characters. A document structure can be considered as a fractal structure.

At the lower abstraction level of a document, more specific information can be obtained. Although a document is not a true mathematical fractal object since a document cannot be viewed in an infinite abstraction level, we may consider a document as prefractal. The smallest unit in a document is character; however, neither a character nor a word will convey any meaningful information concerning the overall content of a document. The lowest abstraction level in our consideration is a term.

The fractal summarization model applies a similar technique as fractal view and fractal image compression (Barnslet, 1988; Jacquin, 1993). An image is regularly segmented into sets of non-overlapping square blocks, called range blocks, and then each range block is subdivided into sub-range blocks, until a contractive mapping can be found to represent this sub-range block. The fractal summarization model generates the summary by a simple recursive deterministic algorithm based on the iterated representation of a document. The original document is represented as a fractal tree structure according to its document struc-

Figure 8. An example of a fractal summarization model



Algorithm 1. Fractal summarization algorithm

```

1. Choose a Compression Ratio.
2. Choose a Threshold Value.
3. Calculate the Sentence Number Quota of the summary.
4. Divide the document into range blocks.
5. Transform the document into fractal tree.
6. Set the current node to the root of the fractal tree.
7. Repeat
7.1 For each child node under current node,
    Calculate the fractal value of child node.
7.2 Allocate Quota to child nodes in proportion to fractal values.
7.3 For each child nodes,
    If the quota is less than threshold value
        Select the sentences in the range block by extraction
    Else
        Set the current node to the child node
        Repeat Step 7.1, 7.2, 7.3
8. Until all the child nodes under current node are processed

```

ture. The system first calculates the sentence significance score based on the summarization features for each sentence. After that, it computes the sum of the normalized sentence significance score of all the sentences under each range block as its Range-Block Significance Score (*RBSS*). The fractal value (*Fv*) of range-block *r* is computed based on the *RBSS* as follows:

$$Fv(r) = \begin{cases} 1 & \text{if } r \text{ is root} \\ C \cdot Fv(\text{parent of } r) \times \left(\frac{RBSS(r)}{\sum_{x \in \text{sibling of } r} RBSS(x)} \right)^{\frac{1}{D}} & \text{otherwise} \end{cases}$$

Given a document, a user will specify compression ratio to specify the amount of informa-

tion displayed. The summarization system calculates the number of sentences to be extracted as summary accordingly, and the system assigns the number of sentences to the root as the quota of sentences. The quota of sentences is allocated to child nodes by propagation—that is, the quota of parent node is shared by its child nodes directly proportional to the fractal value of the child nodes. The quota is then iteratively allocated to child nodes of child nodes until the quota allocated is less than a threshold value and the range block can be transformed to some key sentences by traditional summarization methods.

Figure 8 demonstrates an example of a fractal summarization model. The detail of the algorithm is shown as Algorithm 1.

The *compression ratio* of summarization is defined as the ratio of number of sentences in the summary to the number of sentences in the source document. It was chosen as 25% in most literature because it has been proven that extraction of 20% of sentences can be as informative as the full text of the source document (Morris, Kasper, & Adams, 1992); those summarization systems can achieve up to a 96% precision (Edmundson, 1969; Kepiec et al., 1995; Teufel & Moens, 1997). However, Teufel and Moens (1998) pointed out that high-compression ratio abstracting is more useful, and 49.6% of precision is reported at 4% compression ratio. In order to minimize the bandwidth requirement and reduce the pressure on computing power of mobile devices, the default value of compression ratio is chosen as 4%. By the definition of compression ratio, the sentence quota of the summary can be calculated by the number of sentences in the source document times the compression ratio.

A threshold value is the maximum number of sentences that can be extracted from a range block, if the quota is larger than the threshold value, and the range block must be divided into sub-range block. Document summarization is different from image compression: more than one attractor can be chosen in one range block. It is proven that in the summarization by extraction of a fixed number of sentences, the optimal length of summary is three to five sentences (Goldstein, Kantrowitz, Mittal, & Carbonell, 1999). The default value of threshold is chosen as 5 in our system.

Summarization Features in Fractal Summarization

The weights of sentences under a range block are calculated by the traditional summarization methods described in the former section. However, the traditional summarization features

cannot fully utilize the fractal model of a document. In traditional summarization mode, the sentence weight is static through the whole summarization process, but the sentence weight should depend on the abstract level at which the document is currently viewing at, and we will show how the summarization features can integrate with the fractal structure of a document.

Thematic Feature in Fractal Summarization

Among the thematic features proposed previously, the *tfidf* score of a keyword is the most widely used approach; however, in the traditional summarization, it does not take into account the document structure, therefore modification of the *tfidf* formulation is derived to capture the document structure and reflect the significance of a term within a range block.

The *tfidf* score of term t_i is calculated as followed:

$$w_{ij} = tf_{ij} \log_2 \left(\frac{N}{n} |t_i| \right)$$

where w_{ij} is the weights of term t_i in document d_j , tf_{ij} is the frequency of term t_i in document d_j , N is the number of documents in the corpus, n is the number of documents in the corpus in which term t_i occurs, and $|t_i|$ is the length of the term t_i .

Many researchers assume that the weight of a term remains the same over the entire document. However, Hearst (1993) thinks that a term should carry a different weight in a different location of a full-length document. For example, a term appears in chapter A once and appears in chapter B many times; the term is obviously more important in chapter B than in chapter A. This idea can be extended to other document levels: if you look at the document level, a specific term inside a document should carry the same weight; but if you look at a

Table 1. *tfidf* score of the term “Hong Kong” at different document levels

	Term frequency	Text block frequency	No of Text Block	<i>tfidf</i> Score
Document-Level	1113	1	1	1217.00
Chapter-Level	70	23	23	70.00
Section-Level	69	247	358	105.95
Subsection-Level	16	405	804	31.83
Paragraph-Level	2	787	2626	5.48
Sentence-Level	1	1113	9098	4.03

chapter level, a specific term inside a chapter should carry the same weight, but the a specific term inside two chapters may carry different weights.

As a result, the *tfidf* score should be modified to different document levels instead of the whole document. In the fractal summarization model, the *tfidf* should be defined as term frequency within a range block inversely proportional to frequency of range blocks containing the term; for example:

$$w_{ir} = tf_{ir} \log_2 \left(\frac{N'}{n'} |t_i| \right)$$

Here, w_{ir} is the weights of term t_i in range block r , tf_{ir} is the frequency of term t_i in range block r , N' is the number of range blocks in the corpus, n' is the number of range blocks in the corpus in which term t_i occurs, and $|t_i|$ is the length of the term t_i .

Taking the Hong Kong in the first chapter, first section, first subsection, first paragraph, first sentence of the Hong Kong Annual Report

2000 as an example (see Table 1), the *tfidf* score at different document levels differ significantly; the maximum value is 1217.00 at the document level and the minimum is 4.03 at the sentence level.

Location Feature in Fractal Summarization

Traditional summarization systems assume that the location weight of a sentence is static, where the location weight of a sentence is fixed. However, the fractal summarization model adopts a dynamic approach; the location weight of a sentence depends on which document level one is viewing.

It is known that the significance of a sentence is affected by the location of the sentence inside a document. For example, the sentences at the beginning and the end of a document are usually more important than the others. If we consider the first and second sentences in the same paragraph at the paragraph level, the first

Figure 9. Fractal summarization with location feature only

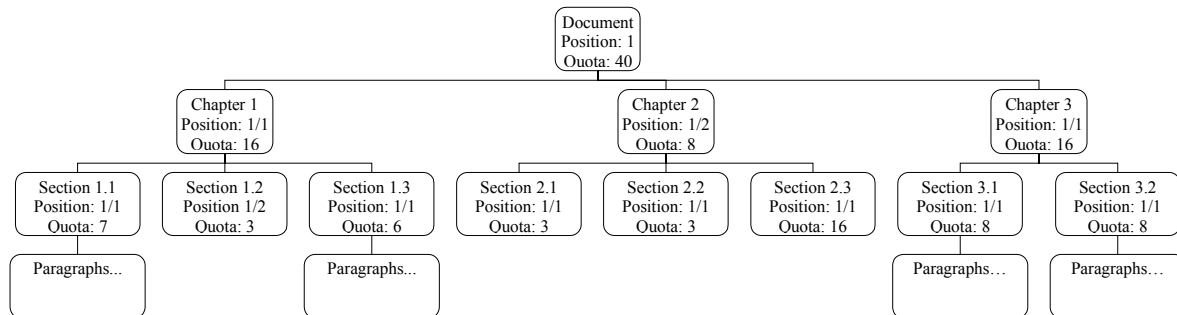
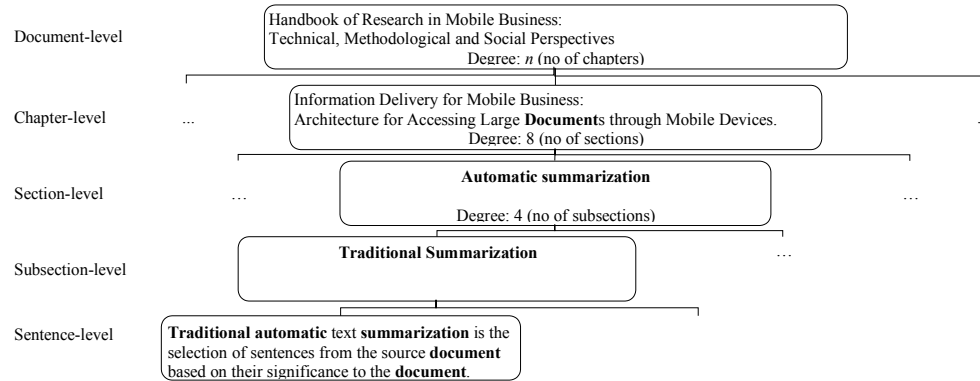


Figure 10. Example of heading feature in fractal summarization



sentence has much more impact on the paragraph than the second sentence. However, the difference of importance of two consecutive sentences is insignificant at the document level. Therefore, the importance of the sentence due to its location should depend on the level we are considering.

In the fractal summarization model, we calculate the location weight for a range block instead of individual sentence; all the sentences within a range block will receive the same location weight. The location weight of a range block is $1/p$, where p is the shortest distance of the range block to the first or last range block under the same parent range block. Consider the previous example of generic fractal summarization model (Figure 8), where the new quota system is changed to Figure 9 if only the location feature is considered.

Heading Feature in Fractal Summarization

During summarization, a sentence containing a term in its headings is considered as more important. The heading weight of a sentence is dynamic and depends on which level we are currently looking at in the document. At a different abstraction level, some headings should

be hidden and some headings must be emphasized.

Taking the first sentence from the first chapter, first section, first subsection, and first paragraph as an example, if we consider the document level, only the document heading should be considered. However, if we consider the chapter level, then we should consider the document heading as well as the chapter heading. Since the main topic of this chapter is represented by the chapter heading, the terms appearing in the chapter heading should have a greater impact on the sentence. Most of the internal nodes above the paragraph level in the document tree usually associate with a heading, and there are two types of headings—structural and informative. The structural headings indicate the structure of the document only, but not any information about the content of the document; for example, “Introduction,” “Overview,” and “Conclusion” are structural headings. The informative headings can give us an abstract of the content of the branch, and they help us to understand the content of the document and are used for calculation of heading weight. On the other hand, the structural headings can be easily isolated by string matching with a dictionary of those structural headings, and they will be used for cue feature.

Table 2. Number of sentences extracted by two summarization models from Hong Kong Annual Report 2000

Chapter ID	Chapter Title	Fractal Summarization	Traditional Summarization
1	Hong Kong: Asia's World City	6	3
2	Constitution and Administration	4	1
3	The Legal System	2	0
4	The Economy	5	14
5	Financial; and Monetary affairs	8	29
6	Commerce and Industry	6	10
7	Employment	2	2
8	Primary Production	1	0
9	Education	2	1
10	Health	1	0
11	Social Welfare	1	0
12	Housing	1	0
13	Land, Public Works and Utilities	4	0
14	Transport	5	3
15	Infrastructure	1	0
16	The Environment	4	1
17	Travel and Tourism	1	1
18	Public Order	5	2
19	Communications, the Media and Information Technology	6	6
20	Religion and Custom	2	0
21	Recreation, Sport and the Arts	5	3
22	Population and Immigration	3	1
23	History	5	3

The terms in the informative headings are very important in extracting the sentences for summarization. Given a sentence in a paragraph, the headings of its corresponding subsection, section, chapter, and document should be considered. The significance of a term in the heading is also affected by the distance between the sentence and the heading in terms of depth in the hierarchical structure of the document. Propagation of fractal value (Koike, 1995) is a promising approach to calculate the heading weight for a sentence.

The first sentence of the subsection “Traditional Summarization” in this chapter is taken as an example to illustrate the propagation of the heading weight (see Figure 10). The heading of this book is “Handbook of Research in Mobile Business: Technical, Methodological, and Social Perspectives.” Assume that there are n chapters, and the heading of this chapter

is “Information Delivery for Mobile Business: Architecture for Accessing Large Documents through Mobile Devices.” The heading of the fourth section in the chapter is “Automatic Summarization.” The heading of the first subsection is “Traditional Summarization.” The first sentence in the subsection is: “Traditional automatic text summarization is the selection of sentences from the source document based on their significance to the document.” To compute the heading weight of the sentence, we shall propagate the term weight of the terms that appear in both the sentence and the headings based on the distance between the headings and the sentences, and the number of text units of the intermediate nodes.

$$\begin{aligned}
 W_{\text{heading}} &= W_{\text{heading in document}} + W_{\text{heading in chapter}} + W_{\text{heading in section}} + W_{\text{heading in subsection}} \\
 \text{where} \\
 W_{\text{heading in document}} &= 0 \\
 W_{\text{heading in chapter}} &= (W_{\text{document}} \times 2) \text{ in chapter heading} / (8 \times 4) \\
 W_{\text{heading in section}} &= (W_{\text{automatic}} + W_{\text{summarization}}) \text{ in section heading} / 4 \\
 W_{\text{heading in subsection}} &= (W_{\text{traditional}} + W_{\text{summarization}}) \text{ in subsection heading}
 \end{aligned}$$

Cue Feature in Fractal Summarization

The abstracting process of human abstractors can help us understand the cue feature at different document levels. When human abstractors extract the sentences from a document, they will follow the document structure to search the topic sentences. During the searching of information, they will pay more attention to the range block with headings containing some bonus words such as “Conclusion,” since they consider it as a more important part in the document and they extract more information for those important parts. The cue feature of the heading sentence is usually classified as the rhetorical feature (Teufel & Moens, 1998).

As a result, we propose to consider the cue feature not only at the sentence level, but also at other document levels. Given a document tree, we will examine the heading of each range block by the method of cue feature and adjust their quota of entire range block accordingly. This procedure can be repeated to sub-range blocks until the sentence level.

Experimental Result

It is believed that a full-length text document contains a set of subtopics (Hearst, 1993), and a good quality summary should cover as many subtopics as possible; the fractal summariza-

tion model will produce a summary with a wider coverage of information subtopics than the traditional summarization model.

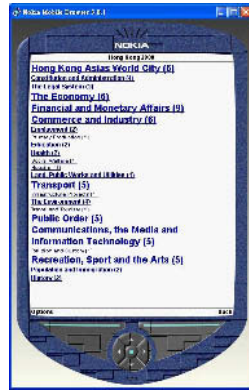
The traditional summarization model extracts most sentences from a few chapters. Using the Hong Kong Annual Report 2000 as an example (see Table 2), the traditional summarization model extracts 29 sentences from one chapter when the sentence quota is 80 sentences, and a total of 53 sentences are extracted from the top three chapters, out of total 23 chapters; not one sentence is extracted from eight of the chapters. However, the fractal summarization model extracts the sentences distributively from each chapter. In our example, it extracts a maximum of eight sentences from one single chapter, and at least one sentence is extracted from each chapter. The standard deviation of sentence number extracted from chapters is 2.11 sentences in fractal summarization vs. 6.55 sentences in traditional summarization. Researchers believe that a good summary should find diverse topic areas in the text and reduce the redundancy of information contents in the summary (Nomoto & Matsumoto, 2001). Fractal summarization extracts the sentences distributively, therefore it finds diverse topic areas and reduces the redundancy of information at the same time.

A user evaluation is conducted. Ten subjects were asked to evaluate the quality of summaries of 23 documents generated by fractal

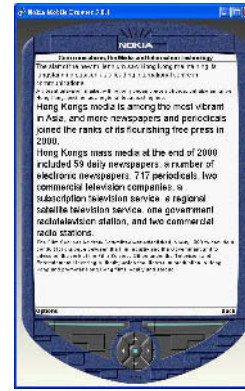
Table 3. Precision of two summarization models

User ID	Fractal Summarization Model	Traditional Summarization Model
User 1	81.25%	71.25%
User 2	85.00%	67.50%
User 3	80.00%	56.25%
User 4	85.00%	63.75%
User 5	88.75%	77.50%
User 6	81.25%	61.25%
User 7	91.25%	76.25%
User 8	86.25%	58.75%
User 9	85.00%	65.00%
User 10	87.50%	72.50%

Figure 11. Screen capture of WAP summarization system



a. Hong Kong Annual Report 2000



b. Chapter 19 of the Hong Kong Annual Report 2000, "Communications, Media and Information Technology"

summarization and traditional summarization. Both summaries of all documents are assigned to each subject in random order without telling the generation methods of the summaries. The results show that all subjects consider the summary generated by fractal summarization method as a better summary. In order to compare the result in greater detail, we calculate the precision as the number of relevant sentences in the summary accepted by the user, divided by the number of sentences in the summary (see Table 3). The fractal summarization can achieve up to 91.25% precision and 87.16% on average, while the traditional summarization can achieve up to a maximum 77.50% precision and 67.00% on average. The one-tailed T-test has shown that the precision of the fractal summarization model outperforms traditional summarization significantly, at a 99% confidence level.

VISUALIZATION OF FRACTAL SUMMARIZATION

The summary generated by the fractal summarization model is represented in a hierarchical

tree structure. The hierarchical structure of summary is suitable for visualization on mobile devices, and it can be further enhanced by displaying the sentences in different font sizes. A summary displayed in a small area without visualization effect is difficult to read. Displaying the sentences in different font sizes according to their importance can help users to focus on important information.

WML is the markup language supported by wireless mobile devices. The basic unit of a WML file is a deck; each deck must contain one or more cards. The card element defines the content displayed to users, and the card cannot be nested. Each card links to another card within or across decks. Nodes on the fractal tree of the fractal summarization model are converted into cards, and anchor links are utilized to implement the tree structure. Given a card of a summary node, there may be a lot of sentences or child nodes. A large number of sentences in a small display area make them difficult to read. In our system, the sentences are displayed in different font sizes according to their significance. We have implemented the system with a three-scale font mode available for WML. The sentences or child nodes are

sorted by their sentence weights or fractal value and separated evenly into three groups. The group with highest value is displayed in “Large” font size, and the group with middle value and the group with lowest value are displayed in “Normal” and “Small” font sizes respectively.

The prototype system using Nokia Handset Simulator is presented in Figure 11. The document we are using is the Hong Kong Annual Report 2000. There are a total of 23 chapters in the annual report, eight of them are in large font, which means that they are more important; the rest are in normal font or small font according to their importance to the report (Figure 11a). The number inside the parentheses indicates the number of sentences under the node that are extracted as part of the summary.

The main screen of the Hong Kong Annual Report 2000 gives the user a general idea of overall information content and the importance of each chapter. If the user wants to explore a particular node, he or she can click the anchor link and the mobile device sends the request to the WAP gateway; the gateway then decides whether to deliver another menu or the summary of the node to the user depending on its fractal value and quota allocated. Figure 11b shows the summary of Chapter 19 of the Hong Kong Annual Report 2000, “Communication, the Media, and Information Technology.”

A handheld PDA is usually equipped with more memory, and the complete summary can be downloaded as a single WML file to the PDA through local synchronization. To read the summary, the PDA is required to install a standard WML file reader.

FINANCIAL NEWS DELIVERY ON MOBILE DEVICES

The fractal summarization model summarizes the documents based on hierarchical document

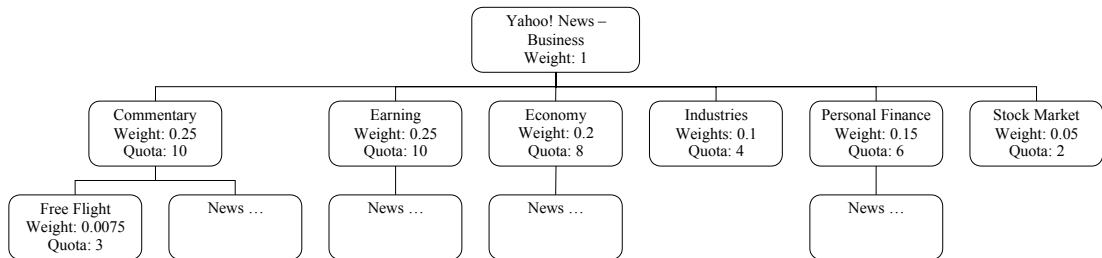
structure. In addition to a large text document, many other information sources also exhibit hierarchical document structure. Due to the large amount of information available, it is difficult to browse these information sources on mobile devices. Automatic summarization is a possible solution. Theoretically, fractal summarization is capable of summarizing all of these information sources, as long as the calculation of fractal value is well formulated. As financial news is critical in decision making, we shall modify the fractal value formula of generic fractal summarization in order to summarize the financial news, and we shall demonstrate financial news delivery with fractal summarization on mobile devices.

Fractal Summarization of Financial News

The fractal summarization model performs summarization based on hierarchical document structure. In addition to large text documents, a lot of other documents also exhibit hierarchical tree document structure, such as Web sites and newspapers. The fractal summarization model is capable of summarizing these documents based on their structure and their relationships in categorization; therefore it is a powerful tool in providing m-services of real-time information delivery. At present, many electronic news delivery services have been provided. An example of the fractal summarization model being used to summarize the financial news available from the Internet is presented in this section.

A newspaper is one of the documents that exhibits the well-defined hierarchical document structure. At present, there are many electronic news delivery services provided for PC, and most of them provide summarization tools to help the user search information, such as the Lycos Financial Feed System with a summarization system from Diyatech, and YellowBrix with Inxight’s Summarizer. How-

Figure 12. Fractal summarization of Yahoo! News–Business category



ever, summarizers for the PC platform are not adaptable to mobile devices directly. Moreover, the existing commercial summarizers are indeed extracting the first few sentences from the document or using the primitive summarization model without considering the hierarchical structure of documents or the organization of information. Yahoo!News is one of the most popular online content providers. There are 21 categories in Yahoo!News. Moreover, each of the categories will be subdivided into subcategories. Take the Business category as an example (see Figure 12). This category contains financial news and is subdivided into six subcategories, namely, Economy, Stock Markets, Earnings, Personal Finance, Industries, and Commentary. Each subcategory contains about 10 news articles. Each news article is a tree structure by itself. For some longer news articles, there may exist more than one section, each section contains a few paragraphs, and each paragraph contains sentences.

The fractal summarization of Yahoo!News is very similar to the fractal summarization of a large text document; only some minor modifications are required to demonstrate the characteristics of Yahoo!News.

- First, the headings of categories and subcategories do not have a direct impact on the content of news under the branch; it

serves for classification purpose only. As a result, the heading method will consider the headings of news articles only. In addition, the headings of categories can be used for personalization of news delivery, the user can set his preference of each category in advance, and the system will adjust the weights accordingly. Alternatively, the preference can be constructed by auto-learning of a machine in the middle tier. The WAP gateway can analyze the reading behavior of the user and predict the user's preference.

- The location feature in traditional summarization assumes that the text unit in the beginning or ending is more important. The news articles inside a subcategory are sorted in chronological order. The most recent news is usually considered as more important. Therefore, we propose calculating the location weight of a news article by its chronological position in the subcategory or the time lag between the news event and browsing time. However, when the system traces the summarization tree down to a node inside a news article, the generic location method in fractal summarization will be adopted.
- In order to provide a glimpse of every article, each news article will receive a sentence quota with at least one sentence.

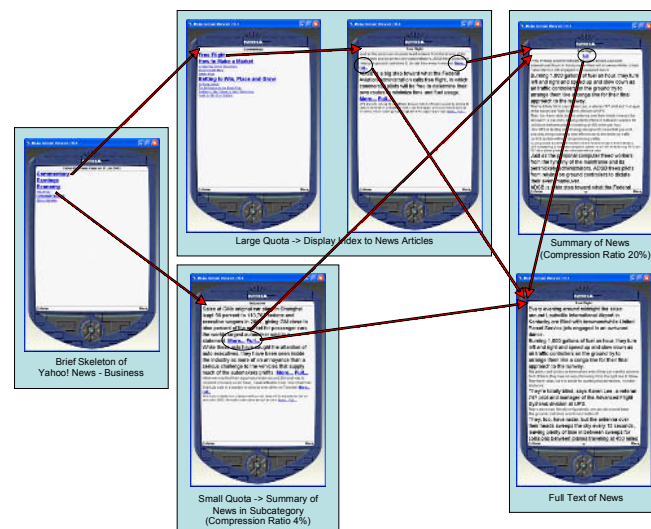
Financial News Delivery to Mobile Devices

In order to minimize the bandwidth requirement and reduce the pressure on computing power of mobile devices, the summarization of Yahoo!News will be conducted in two levels. As high-compression ratio abstracting is more useful (Teufel & Moens, 1998) and can save network bandwidth, the fractal summarization system generates a brief skeleton of summary, with compression ratio equal to 4% at the first stage. The details of the summary at different levels of the news tree are generated on demand by users.

When the mobile device retrieves the financial news from Yahoo!News-Business, the system will first show a card containing six subcategories of the Business category (see Figure 13). In the Figure, three subcategories of the Business category are displayed in large font, which means that they are more important; the rest are in a normal or small font according to their importance. The skeleton of

news gives the user a general idea about how the news articles are organized, and the user can decide in which subcategory to go into detail. When the user clicks the anchor link of subcategory, the WAP gateway will deliver a card depending on the quota allocated. If a large quota is allocated to the subcategory, the system will show another card containing an index of news article. However, if the quota is less than a threshold value of five sentences, the system will show a card with the summary of all news articles in the subcategory. In the summary page, when the user clicks the anchor link 'More' at the end of sentences, the system will generate the summary for the corresponding news articles with a compression ratio of 20%, because it has been proven that the extraction of 20% sentences can be as informative as the full text of the source document (Morris et al., 1992). On the other hand, the user can click the anchor link 'Full' to view the full text of the news articles. Such interactive summarization reduces the computation load, when comparing it with the generation of the

Figure 13. Financial news delivery system on mobile devices



entire summary in one batch by the traditional automatic summarization, which is ideal for m-services.

CONCLUSION AND FUTURE DIRECTION

Mobile business is a promising addition to the electronic commerce by the adoption of portable mobile devices. However, mobile computing should not be limited to user-centered m-service applications only; it should be extended to decision making in an organization. With a fast-paced economy, organizations need to make decisions as fast as possible, and access to large text documents or other information sources is important during decision making. Unfortunately, there are many shortcomings of the mobile devices, such as limited resolution and narrow bandwidth. In order to overcome the shortcomings, fractal summarization and information visualization are proposed in this chapter; these are critical in decision support in an m-organization. Fractal summarization creates a summary in the hierarchical tree structure and presents the summary to the mobile devices through cards in WML. The adoption of keyword feature, location feature, heading feature, and cue feature are discussed. Users may browse the selected summary by clicking the anchor links from the highest abstraction level to the lowest abstraction level. Based on the sentence weight computed by the summarization technique, the sentences are displayed in different font size to enlarge the focus of interest and diminish the less significant sentences. Such visualization effect draws users' attention to the important content. The three-tier architecture is presented to reduce the computing load of the mobile devices. The proposed system creates an information visualization environment to avoid the existing shortcomings of mobile devices for mobile business.

In its current stage, fractal summarization is capable of processing textual information only. However, there is a lot of information available in multimedia formats on the Web. Information delivery of multimedia documents will be one of key research topics in the near future. As multimedia documents require a much higher bandwidth than textual documents, this problem cannot be resolved solely by the current steaming technology. Summarization of multimedia documents is required for information delivery to mobile devices. The research of spoken document summarization and video summarization has been started (Vasconcelos & Lippman, 1998; Zechner & Waibel, 2000). It would be a great challenge to move the proposed model to multimedia documents. The summarization of multimedia documents is complementary to the proposed model. Nowadays, most of the mobile devices are speech based. With the summarization of spoken documents, the information can be easily delivered to speech-based mobile devices. This will certainly increase the popularity of the proposed model.

REFERENCES

- Barnsley, M. F., & Jacquin, A. E. (1988, November). Application of recurrent iterated function systems to images. *Proceedings of the Conference on SPIE Visual Communications and Image Processing (VCIP'88)*, Cambridge, MA (vol. 1001, pp. 122-131).
- Baxendale, P. B. (1958). Machine-made index for technical literature—An experiment. *IBM Journal of Research and Development*, 2(2), 354-361.
- Brown, M. H., & Weihl, W. E. (1996, October). Zippers: A focus + context display of Web pages. *Proceedings of the World Conference of the Web Society (WebNet'96)*, San Francisco, CA.

- Buyukkokten, O., Garcia-Molina, H., & Paepcke, A. (2001a, May). Seeing the whole in parts: Text summarization for Web browsing on handheld devices. *Proceedings of the 10th International Conference on the World Wide Web (WWW10)*, Hong Kong, China (pp. 652-662).
- Buyukkokten, O., Garcia-Molina, H., & Paepcke, A. (2001b, March). Accordion summarization for end-game browsing on PDAs and cellular phones. *Proceedings of the SIGCHI Conference on Human Factors in Computing System (CHI 2001)*, Seattle, WA (pp. 213-220).
- Buyukkokten, O., Garcia-Molina, H., & Paepcke, A. (2001c, June). Text summarization of Web pages on handheld devices. *Proceedings of the Workshop on Automatic Summarization 2001 in conjunction with the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL 2001)*, Pittsburgh, PA.
- Buyukkokten, O., Garcia-Molina, H., Paepcke, A., & Winograd, T. (2000, April). Power browser: Efficient Web browsing for PDAs. *Proceedings of the SIGCHI Conference on Human Factors in Computing System (CHI 2000)*, Hague, The Netherlands (pp. 430-437).
- Edmundson, H. P. (1969). New method in automatic extraction. *Journal of the ACM*, 16(2), 264-285.
- Endres-Niggemeyer, B., Maier, E., & Sigel, A. (1995). How to implement a naturalistic model of abstracting: Four core working steps of an expert abstractor. *Information Processing and Management*, 31(5), 631-674.
- Feder, J. (1988). *Fractals*. New York: Plenum.
- Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory; strategies for qualitative research*. New York: Aldine de Gruyter.
- Goldstein, J., Kantrowitz, M., Mittal, V., & Carbonell, J. (1999, August). Summarizing text documents: Sentence selection and evaluation metrics. *Proceedings of the 22nd Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR '99)*, Berkeley, CA (pp. 121-128).
- Harman, D. K. (1992). Ranking algorithms. In W.B. Frakes & R. Baeza-Yates (Eds.), *Information retrieval: Data structures and algorithms* (pp. 363-392). Englewood Cliffs, NJ: Prentice-Hall.
- Hearst, M. A. (1993, June). Subtopic structuring for full-length document access. *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '93)*, Pittsburgh, PA (pp. 56-68).
- Jacquin, A. E. (1993). Fractal image coding: A review. *IEEE*, 81(10), 1451-1465.
- Kepiec, J., Pedersen, J., & Chen, F. (1995, July). A trainable document summarizer. *Proceedings of the 18th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR '95)*, Seattle, WA (pp. 68-73).
- Koike, H. (1995). Fractal views: A fractal-based method for controlling information display. *ACM Transactions on Information Systems*, 13(3), 305-323.
- Lam-Adesina, M., & Jones, G. J. F. (2001, September). Applying summarization techniques for term selection in relevance feedback. *Proceedings of the 24th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR '01)*, New Orleans, LA (pp. 1-9).
- Lin, Y., & Hovy, E. H. (1997, March). Identifying topics by position. *Proceedings of the Workshop on Intelligent Scalable Text Sum-*

- marization in conjunction with the 5th Conference on Applied Natural Language Processing Conference (ANLP'97), Washington, DC (pp. 283-290).
- Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2), 159-165.
- Mandelbrot, B. (1983). *The fractal geometry of nature*. New York: W. H. Freeman.
- Mani, I. (2001, November). Recent development in text summarization. *Proceedings of the 10th International Conference on Information and Knowledge Management (CIKM'01)*, Atlanta, GA (pp. 529-531).
- Morris, A. H., Kasper, G. M., & Adams, D. A. (1992). The effects and limitations of automated text condensing on reading comprehension performance. *Information System Research*, 3(1), 17-35.
- Nomoto, T., & Matsumoto, Y. (2001, September). A new approach to unsupervised text summarization. *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'01)*, New Orleans, LA (pp. 26-34).
- Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5), 513-523.
- Teufel, S., & Moens, M. (1997, July). Sentence extraction as a classification task. *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent and Scalable Text Summarization*, Madrid, Spain (pp. 58-68).
- Teufel, S., & Moens, M. (1998, March). Sentence extraction and rhetorical classification for flexible abstracts. *Proceedings of the 1998 AAAI Spring Symposium on Intelligent Text Summarization*, Palo Alto, CA (pp. 16-25).
- Vasconcelos, N., & Lippman, A. (1998, June). A spatiotemporal motion model for video summarization. *Proceedings of the IEEE Computer Society Conference on computer Vision and Pattern Recognition (CVPR'98)*, Santa Barbara, CA (pp. 361-366).
- Yang, C. C., & Wang, F. L. (2002, December). Document summarization on handheld device: An information visualization tool for mobile commerce. *Proceedings of the First Workshop on E-Business (WEB2002) of the International Conference on Information Systems (ICIS 2002)*, Barcelona, Spain.
- Yang, C. C., & Wang, F. L. (2003, July). Fractal summarization: Summarization based on fractal theory. *Proceedings of the 26th Annual International ACM SIGIR Conference: Research and Development in Information Retrieval (SIGIR 2003)*, Toronto, Canada (pp. 391-392).
- Yang, C. C., & Wang, F. L. (2003, May). Automatic summarization for financial news delivery on mobile devices. *Proceedings of the 12th International Conference on the World Wide Web (WWW2003)*, Budapest, Hungary (pp. 391-392).
- Yang, C. C., & Wang, F. L. (2003, May). Fractal summarization for mobile devices to access large documents on the Web. *Proceedings of the 12th International Conference on the World Wide Web (WWW2003)*, Budapest, Hungary (pp. 215-224).
- Zechner, K., & Waibel, A. (2000, July). DiaSumm: Flexible summarization of spontaneous dialogues in unrestricted domains. *Proceedings of the 18th International Conference on Computational Linguistics (COLING-2000)*, Saarbruecken, Germany (pp. 968-974).

Chapter XIX

Database Queries in Mobile Environments

N. Marsit

IRIT—Paul Sabatier University, France

A. Hameurlain

IRIT—Paul Sabatier University, France

Z. Mammeri

IRIT—Paul Sabatier University, France

F. Morvan

IRIT—Paul Sabatier University, France

ABSTRACT

The technological evolution of networks together with the development of positioning systems has contributed to the emergence of numerous location-based services. Services related to this expanding area will become of major technical as well as economical interest in the coming few years. This aroused a great deal of interest from the scientific community at large and specifically from those studying these services and their diverse requirements and constraints. One of the direct consequences in the database field is the appearance of new types of queries (mobile queries issued from mobile terminals and/or requesting information associated with moving objects such as vehicles). Our objective in this chapter is to present a comprehensive survey of the field of research work related to mobile queries, with particular attention to the location issue.

INTRODUCTION

Mobile units are obviously on the rise. Thanks to the technological progress realized in this domain, mobile terminals and units have become successful and widely used by the gen-

eral public. At the beginning of the wireless revolution, the main objective of wireless networks was to enable mobile units to communicate. Nowadays, networks support various new services and applications. In fact, the significant technological evolutions added to the de-

velopment of positioning systems such as GPSs have contributed to the emergence of a large number of location-based services and applications (e.g., a mobile user asking for data related to his location such as the closest hotel). These types of applications are about to become the major focus of economical interest in the next few years. The location-based service benefits are expected to exceed \$40 billion in 2006, while they were estimated to approximately \$1 billion in 2000 (Mokhtar & Su, 2004). This has aroused the interest of a great part of the scientific community devoted to research and development in this area. One of the direct consequences in the database field is the appearance of new types of queries. In mobile environments, there are two basic categories of queries (Marsit, Hameurlain, Mammeri, & Morvan, 2005). The first one includes queries issued from mobile terminals and querying data related to fixed objects (such as hotels, gas stations, hospitals)—for example, “select the closest restaurants.” The second category includes the queries issued from mobile or fixed terminals and querying data related to moving objects (such as vehicles, helicopters, boats, people). For example, “select all ambulances that will be at 2 km from the hospital within 10 minutes.” Within these two categories we can distinguish different types of queries according to their location dependence, to the association of spatial and temporal dimensions and to the evaluation mode (continuous or not).

The main objective of this chapter is to review work related to mobile queries (i.e., queries issued by mobile terminals and/or querying data related to moving objects). We start by classifying the different types of mobile queries. This step is central because it allows us to highlight the constraints of each type of query and to identify their underlying problems. In the field of mobility, various topics and problems were addressed by several different research communities. We only present the

problems generated by the requirements of the new types of mobile queries. Finally, we point out the still open problems and identify the new challenges related to query processing in mobile environments.

QUERY CLASSIFICATION

Context

In mobile environments, entities can be either fixed or mobile. Hence, defining what mobility means is an essential requirement:

- **Mobile Client:** The query is submitted by a mobile terminal, here called mobile client.
- **Mobile Server:** The query or part of the query is processed at one or several mobile servers.
- **Moving Object:** Data targeted by the query can represent, in databases, moving objects (e.g., vehicles).

In this context we can highlight some query classification criteria. The first is the mobility constraint which allows distinguishing two basic forms of queries: (1) queries submitted by mobile terminals and querying data related to fixed objects (e.g., hotels), and (2) queries submitted by mobile or fixed terminals and querying data related to moving objects.

Notice that the mobility of servers does not add additional types of queries. Nevertheless, it may have an impact on query execution models since other problems have to be considered (e.g., network connection, server localization, etc.) (Holliday, Agrawal, & Abbadi, 2002).

For both query categories mentioned above, other types of queries could be distinguished. In fact, a second criterion, location constraint, brings out three types of queries: *Non-Loca-*

tion-Related Query, *Location-Aware Query*, and *Location-Dependent Query*. The third criterion depends on the association of the spatial and temporal dimensions. Finally, according to the query evaluation mode, we distinguish another type called continuous queries.

Location-Dependent Queries

Seydim, Dunham, and Kumar (2001a) present a framework for the location relatedness in queries. They consider that attributes, relations, operators, and simple predicates can be either location-related or not location-related:

- An attribute is considered as location-related if its domain is location-related. For instance, in the relation *Hotel* (*id-Hotel*, *name*, *occupancy*, *city*, *street*), the attributes *city* and *street* are location-related attributes.
- A location-related relation contains at least one location-related attribute. Otherwise, it is considered a non-location-related relation. (e.g., the relation *Hotel* is location-related because it contains the location-related attributes *city* and *street*).
- An operator is location-related if at least one of its operands is location-related.
- A simple location-related predicate is a simple predicate where the operator is location-related and the operands are defined in location-related domains.

Based on these definitions we can describe the types of queries mentioned above.

Non-Location-Related Queries (NLRQs)

If all the predicates and attributes used in a query are non-location-related, then it is called

a non-location-related query (Seydim et al., 2001a)—for example, *select availability of hotel with identifier 10*. NLRQs are considered in many cases as traditional queries. However, the issuer of the query is mobile. So, peculiar problems specific to mobility like location management have to be addressed. These problems are discussed later.

Location-Aware Queries (LAQs)

If a query has at least one location-related simple predicate or one location-related attribute, then it is called Location-Aware Query—for example, *select the names and the availabilities of the hotels in Toulouse*. To express such a query, we need the attribute *city*, which is location-related. This is sufficient to consider the query as LAQ. Other examples may highlight the need for a new special location-related operator such as “*close to*” to express proximity—for example, *select restaurants, close to the hotel, whose identifier is 10*. Thus, we underline the need to define new methods to evaluate this type of operator. Moreover, several studies considered the problem of defining new location-related operators and efficient methods to process them (Seydim et al., 2001a).

Location-Dependent Queries (LDQs)

If the query results depend on the location of the query issuer (the mobile client), then the query is called a Location-Dependent Query—for example, *select the closest hospital*. Note that LDQ processing brings new challenges to the database community. First, new operators that take into account the notion of proximity (*close*, *closest to*) and orientation of the mobile client (*straight ahead*) have to be defined. Second, another step is required to bind the location of the mobile client to the query. The query issuer location could be recovered from

different sources (wireless network operator's databases, positioning system such as GPS, etc.). Another problem occurs when the mobile client location is determined with a granularity that differs from the one of the data stored in the database. Indeed, a process of changing the granularity of the mobile client location given by the location service to the appropriate granularity required by the application has to be considered.

Moving Object Database Queries (MODQs)

This type includes queries issued by mobile or fixed terminals and querying moving object databases (i.e., databases in which data represent moving objects such as vehicles or planes). This type of query has appeared with the emergence of numerous applications requiring moving object data storage capability (Seydim et al., 2001a; Sistla, Wolfson, Chamberlain, & Dao, 1997). Such applications owe their popularity to the increasing development of positioning systems and technologies, allowing real-time tracking of moving objects—for example, *select all taxis that are now in Wilson Square*. For such a query, the size and the shape of the object are not important. Generally, the position of the moving object is the most requested information. Hence, several problems arise such as modeling and querying moving objects with rapidly changing locations, tracking and updating the location of moving objects, and managing the uncertainty on location due to imprecision of sensor technology and to the continuous movement of moving objects.

Spatio-Temporal Queries

The spatio-temporal type includes queries combining space dimension with time dimension (Erwig, Güting, Schneider, & Vazirgiannis, 1999). In the literature, this type of query is often associated with moving objects. Indeed,

numerous applications do not only focus on the location of an object, but also on its position at a given time or on its trajectory during a certain time interval. The time notion generally implies the past, the present, and the future. Thus, we can distinguish between two types of spatio-temporal queries:

- The first type considers trajectories describing a time history of object movement—for example, *select all moving objects with trajectories included within the area 'R' between 04:00 p.m. and 05:00 p.m.*
- The second type focuses on the current position of the moving object and possibly its future position—for example, *select all ambulances that will be 2 km away from the hospital in less than 10 min.*

The spatio-temporal queries raise problems at several levels. Representing continuously changing data (e.g., position) has been one of the major points of interest during the past few years. Complex structures have to be managed to represent moving object trajectories. In addition, effective methods must be established to represent the moving object movement and to predict its future positions. The extension of traditional languages, such as SQL, enables support of querying spatial or temporal data. However, they have to be further extended to really represent the strong relationship between space and time. Finally, the notion of uncertainty, whether in the data model or in the proposed extensions of languages, has to be dealt with. We present in the fourth section different proposals developed in this research area.

Continuous Queries

A continuous query (CQ) allows users to get changing results from a database without hav-

Table 1. Summary of the mobile query types

Query type	Mobile client	Moving object	Spatial dimension	Temporal dimension
NLRQ (Non-Location-Related Query)	Yes	No	No	No
LAQ (Location-Aware Query)	Yes	No	Location (explicit)	No
LDQ (Location-Dependent Query)	Yes	No	Location of client (implicit)	No
MODQ (Moving Object Database Query)	Yes/No	Yes	Yes	No
Spatio-Temporal Query	Yes/No	Yes	Yes	Yes
CQ (Continuous Query)	Yes/No	Yes/No	Yes/No	Yes/No

ing to issue the query repeatedly (Chen, Dewitt, Tian, & Wang, 2000). Assume that a driver is asking for a selection of hotels within 5 km from his position. If this query is submitted as non-continuous, then the results are sent back immediately after the query has been processed. If the same query is submitted as continuous, then the set of selected hotels varies “continuously” with the movement of the user. Notice that continuous queries require considerable modifications in the query evaluation algorithms. In fact, issues such as “when” and “how often” the continuous query should be re-evaluated have to be addressed. Also, the possibility of partial or incremental re-evaluation has to be investigated (Gök & Ulusoy, 2000).

Discussion

In this section, we presented the types of query which appear to be most crucial in mobile environments. However, we must insist here on two essential points: first, certain types of queries previously presented form non-separate groups. LDQ, MODQ, and spatio-temporal queries can be submitted as continuous queries. Nevertheless, with our classification we covered the types of mobile queries most often encountered in the literature, and we focused

on their characteristics. In Table 1, we summarize six types of queries presented in this chapter. We observe that the first three types (NLRQ, LAQ, and LDQ) involve only the mobility of the client, whereas the other types involve moving objects and/or the mobility of clients. The fundamental difference between LDQ and LAQ is that for LDQ the notion of location is implicitly involved, while it is explicitly specified for LAQ. So, the LDQ type can be considered as part of the LAQ type. Spatio-temporal queries add a temporal dimension to the spatial dimension. Finally, note that CQ is a transverse type which can be associated to the other types of queries.

A second essential point concerns the common problems met for the various types of queries. Indeed, certain problems can be recurring in all query types (e.g., the management of mobile localization). Then, certain types of queries share very close problems such as data modeling of moving objects and representation of spatio-temporal data on moving objects. Moreover, considerable effort was devoted to model moving objects in the final objective to support spatio-temporal queries (Forlizzi, Güting, Nardelli, & Schneider, 2000; Grumbach, Rigaux, & Segoun, 2000; Sistla et al., 1997).

QUERYING FIXED-OBJECT DATABASES

The mobility domain is very large, and it would be over ambitions to present an exhaustive state of the art on all work related to this field. We rather focus on the problems related to query processing in mobile environments. At the beginning of the previous section, we classified the mobile queries into two basic categories: those which are issued from mobile terminals and which are querying data on fixed objects, and those which are querying data on moving objects. In this section we describe problems of the first query class and give some methods proposed to cope with them. In the next section, we consider related work on the second category of queries.

Localization of Mobile Units

Since the beginning of the 1990s, localization of mobile units is one of the major concerns of researchers interested in mobility. In fact, while a mobile user submits a query, the system should be able to locate him in order to send back the answer. For these reasons, the localization of mobile units is an important issue related to query processing in mobile environments.

Location Models

The location model for mobile units is closely dependent on sensor systems used to detect the location. There are two basic models for representing the location: symbolic models and geometric models (Lee, Lee, Xu, & Zheng, 2002). The use of symbolic or geometric models depends on the application to be developed because they meet different requirements in terms of location representation and required precision.

In the symbolic model, the location is represented by entities from the real world such as streets, cities, and zip code. In this model, location can also be represented by elements defined in any particular systems, for instance, cells in a cellular system. The location information in a symbolic model is well structured and easy to manage. Their granularity is well suited for several location-based applications and services because their representation is often based on relationships between entities (e.g., a street is in a city, a city is in a zip code). However, this granularity is not very fine and it often depends on applications or on used systems.

In the geometric model, the location is represented by n-dimensional coordinates (generally 2 or 3). This model can give good accuracy and is compatible with heterogeneous systems. However, it can be costly in terms of data volume and sometimes the location information needs to be translated into a level understandable by the application.

Location Management

It is important to know the current location of mobile units. This information is usually stored at specific network sites. The main issue here is to find a compromise between the update cost and the lookup cost of location information. In fact, to reduce the cost of lookup, it is necessary to increase the number of sites where this information is stored. Hence, the availability of the location information is improved but the cost of updates becomes higher. On the other hand, if the frequency of the update is reduced, then the precision of the location information is compromised. In this area, various strategies that balance the cost of lookups against the cost of updates were proposed. The main relevant approaches are based on two types of location database architectures: two-tier schemes and hierarchic schemes. In Pitoura

and Samaras (2001), a comprehensive survey of locating objects in mobile environments is proposed.

Location-Dependent Query Processing

The proposed techniques of localization were generally dedicated to wireless network operators. They allow mobile users to communicate or to get the answer of a query. With the development of location-based services and applications, several new requirements arose in terms of data querying. In fact, the appearance of new types of queries like LDQ changed the problem of localization, since the location of a mobile client is involved even in query processing. This led to much research on LDQ processing.

Binding Location to LDQ

To better analyze the problem of LDQ query processing, let us take the example: *select the closest hospital*. The idea presented in Seydim et al. (2001a, 2001b) is to translate the query into *select the closest hospital to my current position*. A question arises here: how to bind the mobile client location to query, especially when this information must be given by network operator databases or by particular sensor network technologies such as GPSs? The proposed solution is based on the use of a particular service called *location service*. A mobile client identifier is sent to the location service which returns its actual location (Leonhardi & Kubach, 1999; Seydim et al., 2001a, 2001b). This service allows developing applications regardless of operator or sensor system. Location information is generally gathered from different sources: positioning system (e.g., GPS, GSM) and network operator databases. Thus, with the help of location service, our query

becomes *select hospitals closest to position X* (where X is the value corresponding to the mobile client location sent back by the location service). This step is called *location binding* (Seydim et al., 2001a).

Matching Location Granularity Level

The location granularity sent by the location service and the location granularity required by the application may not be the same. Assume that the position X of the example studied above is sent according to the geometric model (e.g., latitude/longitude) and the hospital locations are represented in the database by zip code or by city name. In this case, a process of translating the query location granularity into the appropriate location granularity as needed by the application is necessary. This process is called *location leveling* (Seydim et al., 2001b). The problem of granularity mismatch may persist even when a unique location model is used. For example, cell and city (symbolic model) do not have the same level of granularity. The first solutions proposed to solve this problem were based on a location translation mechanism (ESRI, 2000; SignalSoft, 2000). These assumed that the set of location granularity needed by applications and those needed by queries are known in advance. So, translations are well defined and explicitly specified by mapping functions. Seydim et al. (2001b) proposed a general process leading to the coordination of dynamic location granularity levels. This process is based on metadata describing the hierarchy and relationship between different location granularities possibly used by applications and queries. Assume for example that the given position X is cell number 3. While consulting the hierarchy stored in metadata, one can find that the cell number 3 is in the zip code 31000. Thus, our query is finally transformed into *select hospital within zip code 31000*.

Processing of Location-Related Operators

Another aspect to which many researchers paid particular attention concerns processing of location-related operators. Indeed, the location constraint in queries implies necessarily the use of this type of operators. Although spatial operators (e.g., *Distance*, *Intersect*, *Contains*, *Within*) form an important part of location-related operators, they are not sufficient to deal with all the requirements of LDQ. So, new operators have to be introduced to express proximity (e.g., *closest to*) and mobile client orientation (e.g., *straight ahead*). Some of these operators can have different semantic interpretations according to the application requirements. In Seydim et al. (2001a), the operators *closest to* and *straight ahead* are treated in spatial manner. Thus, for the *closest to* operator, the authors define an area around the query issuer to access the related data. This area may be a circle, a half circle, and so on. For the *straight ahead* operator, they define a window to select an area ahead of the direction of the user. These operators are not always interpreted in spatial manner. In fact, *closest to* may mean *object from the same city or the same zip code or the same cell or neighbor cell*.

To close this section, let us recall that in this chapter we were only interested in two aspects (i.e., localization of mobile terminals and query processing) of work related to processing of queries issued by mobile terminals. Obviously, other research directions were developed to solve problems of cache coherency, of transaction management (Serrano-Alvarado, Roncancio, & Adiba, 2001), and of data access in mobile environments (Birman et al., 1999).

QUERYING MOVING-OBJECT DATABASES

Much effort has been devoted to solving problems related to processing queries issued by mobile terminals and querying data related to fixed objects and more particularly LDQ. However, fast evolution of technologies providing increasingly precise information on location and movement of objects, like GPSs, encouraged the development of new types of applications. Indeed, the past few years have witnessed the emergence of a wide range of complex applications managing moving objects (e.g., fleet management, air traffic management, road traffic management, emergencies). These new applications have generated new constraints and new requirements. Hence, the database community is facing new challenges. In this section, we expose a part of the most representative topics of research related to querying moving-object databases. We concentrate on the problems of modeling, querying, language extensions, and uncertainty management. Obviously, other problems were dealt with in the literature such as indexing spatio-temporal data. Although these aspects are also important, we will not present them in this chapter.

Modeling of Moving Objects

Representing the continuous movement of objects in databases is one of the central problems of modeling moving objects. First, the computer systems are not able to handle infinite sets. Indeed, classical DBMSs consider data unchanged until they are explicitly updated. This is not sufficient to represent continuously changing information such as the position of a moving object. Second, positioning systems return locations of objects in a discrete manner. This

problem is similar to the one of modeling spatio-temporal data. In fact, a spatio-temporal object is defined as an object whose shape or position varies as a function of time. The shape of moving objects is often ignored since they are generally assimilated to moving points. This assumption simplifies the problem compared to the more general one of representing spatio-temporal data. Approaches proposed to represent the spatio-temporal behavior of moving objects follow two main directions: the first centers on representing a history of the movement and trajectories of the moving objects; the second focuses on modeling the current location of a moving object and its possible future position. In the first direction, we distinguish two approaches: one which proposes to extend the existing systems with abstract data types, while the second is based on constraint databases. The second direction relies on a dynamic attributes modeling approach.

Abstract Data Types

The idea of extending the classical DBMS features by Abstract Data Type (ADT) appeared near the end of the 1980s (Güting, 1989). Nearly 10 years later, Erwig et al., (1999) suggested exploiting the ADT concept to define spatio-temporal ADT. This work was pursued with the aim of introducing a system of ADT with suitable operations into the DBMS. This way, it is possible to better represent spatio-temporal data and to extend query languages (Forlizzi et al., 2000; Güting et al., 2000). Erwig et al. (1999) conclude that two levels of abstraction are necessary. The first level, called abstract model, is relatively simple (Güting et al., 2000). It allows handling infinite sets without considering details of their representation. However, the implementation of this model is not obvious. The second level (a discrete model) makes it possible to implement

the types introduced in the abstract model by associating each data type of the abstract model with “discrete” types whose domains are defined according to a finite representation.

The abstract model relies on base types (*int*, *real*, *string*, *bool*), spatial types of 2D dimension (*point*, *points*, *line*), and temporal types (*intime*). One of the most important types of constructors is *moving*. It enables construction of types whose values change dynamically. For example, a *moving(point)* value is a function defined from time into point values. So, with this constructor one can represent the movement of objects. One can also represent a moving region by *moving (region)*. Moreover, several operations on these types can be defined such as *at*, *mdistance*, *trajectory*, and so forth.

Further comprehensive explanations can be found in several papers (Erwig et al., 1999; Forlizzi et al., 2000; Güting et al., 2000).

Constraint Database Model

The constraint database model was initially proposed by Kanellakis, Kuper, and Revesz (1990, 1995). Even if the paradigm presented allows the representation of any type of data, this idea especially seduced the spatial database community. The principle is to consider the spatial objects as infinite sets of points satisfying first-order logic formulae (these formulae correspond to constraints). The constraint representation can be viewed as an extension of finite relational representations. In fact, a relation is considered as a first-order logic formula applied to infinite sets. Grumbach et al. (2000) relied on constraint database to represent and manipulate multidimensional data. They illustrated their model with spatio-temporal applications manipulating time and spatial objects. They considered space and time as natural components of 3D point sets. Thus, the spatio-temporal data are seen as mathematical

objects which can be analyzed with already known tools (Mokhtar, Su, & Ibarra, 2002). So the location of moving objects can be modeled by a linear function mapping time to n -dimensional space. A trajectory is generally approximated by a succession of segments connected pairwise in a three-dimensional space. Each segment is represented as a conjunction of linear constraints using time variables and coordinate variables. The complete trajectory is represented by the disjunction of all its linear constraints.

Let us retain that the abstract representation of this model is completely independent of the implementation and query processing algorithms. In addition to this model, query processing has a reasonable complexity and query languages are simple. In fact, users can use standard languages because they do not have to care about internal mechanisms (data structures, operations) (Grumbach et al., 2000). However, this model is not suited for spatio-temporal queries which involve future locations of objects. Little work tried to extend this model in order to support this type of query.

Model Based on Dynamic Attributes

Sistla et al. (1997) proposed a model called MOST (Moving Objects Spatio-Temporal). This model was designed to be implemented in a software layer on the top of existing DBMSs. A prototype was performed within the framework of the DOMINO project (Wolfson, Sistla, Xu, Zhou, & Chamberlain, 1999b). The objective of this work is to extend classical DBMS features in order to represent in an effective way the current location of moving objects and to enable prediction of their future positions. The originality of this work lies in the introduction of the notion of dynamic attributes. The values of dynamic attributes change over time according to a given function without having to

explicitly update them. Thus, within the MOST model, object attributes can be either static or dynamic. Let us take the object airplane as example. It includes static attributes such as the identifier and the company name, dynamic attributes such as the position represented by coordinate (X, Y, Z) . Each dynamic attribute is represented by three sub-attributes. For example X is represented by $X.value$ (value of X at the last update time), $X.updatetime$ (the last update time), $X.function$ (function of a variable t . at $t=0$, $X.function=0$). This function has to represent the motion vector of the object (e.g., speed). At each time $t+updatetime$, the value of X is given by computing $X.value+X.function(t)$. Hence, with the MOST model, we can implicitly deduce the future positions of moving objects and answer queries which concern the future state of the database. In Sistla et al. (1997), the FTL (Future Temporal Logic) language is proposed in order to express this type of query.

Languages for Querying Moving Object Databases

The expression of moving objects database queries is closely tied to the chosen model. Therefore, several proposals of language extensions were introduced according to the underlying model of data. In the following, we will focus on language extensions proposed for each approach.

Querying in Abstract Data Type Model

Within the abstract model, semantics of operations and functions are defined in a relatively simple way. Indeed, they can handle infinite sets without considering their physical representation. Many operators are defined in Forlizzi et al. (2000) and Güting et al. (2000). These operators can be incorporated into languages

intended for users (e.g., extended SQL). To illustrate this, let us take the following example:

airplane (id : string, company : string, pos : mpoint) is a relation. mpoint is moving(point). pos is a spatio-temporal attribute, which represents the location of the plane as a function of time. Assume that we have two operators, trajectory and length, defined as follows:

trajectory: mpoint $\rightarrow$ line, returns a line which is the projection of moving point in plan.

Length: line $\rightarrow$ real, determines the line length.

Now, we can express the query such as select Air France planes having a trajectory longer than 3000 miles:

```
select id
from airplane
Where company = "Air France" and
length(trajectory(pos)) > 3000;
```

Obviously, in this framework further operators are defined in order to express more spatio-temporal queries.

Querying in Constraint Database Model

The particularity of this model lies in the fact that it hides from users the complexity of the data model. Indeed, this model allows handling infinite sets of points, forming tuples of relations, without caring about their finite representations (managed by the model). Therefore, the user can simply use a standard language like SQL. Assume that the relation trajectory(X, Y, Z) represents the position of an object at any time; the user does not care about the finite representation of this abstract relation (infinite). Thus, to ask a query such as select position of the plane at time t1, we have only to use SQL language as follows:

```
Select X, Y, Z from trajectory where t = t1;
```

Naturally, the evaluation of these queries requires algorithms based on finite representation. Grumbach et al. (2000) propose a method exploiting the interpolated form of data (to obtain one of the variables describing an object as a function of the other variables). The query evaluation is then reduced to a small number of operators applied to a sub-space formed by key attributes (in the constraint model, all the attributes can be expressed as a function of key attributes of the infinite relations). Thus, the selection operation can be evaluated by a simple conjunction formula. Other language extensions and other evaluation methods were proposed in the framework of the constraint database model (Mokhtar et al., 2002).

Querying in MOST Model

Sistla et al. (1997) developed the MOST model, which allows the management of the current position of a mobile object and to anticipate its future movement. In this framework, an extension of standard languages such as SQL or OQL by temporal logic predicates was proposed. This language, called FTL, is based on two basic temporal operators: *Until* and *Nexttime*. *f until g* is satisfied if and only if the predicate *g* is true in this state or *g* is satisfied in the future and until then *f* is satisfied. *Nexttime f* is satisfied if the formulae *f* is satisfied at the next state of the history. From these two basic operators, other temporal operators can be expressed in FTL: *Eventually f* = *true Until f*; *Always f* = $\neg$ *Eventually $\neg$ f*; *Eventually_within_c g* (i.e., *g* will be satisfied within *c* time units).

Given that the language and the model are designed to be implemented on top of DBMS, the FTL language is assumed to support non-temporal operators of the underlying languages. Thus, by means of this language, we can ex-

press the spatio-temporal queries in a simple and intuitive way.

Example: Select all the planes which enter a region P around the airport within 30 min.

Assume the relation airplane (id, company, X_pos). Assume that the X_pos attribute gives the geographical coordinates of the plane and that a polygon P is defined as a spatial object. Assume also that a spatial operator Inside (X_pos, P) returns true if the first argument X_pos is inside a given area P. Else, it returns false. We can now express the query using the temporal operator already defined in FTL:

```
Eventually_within_c:
Select id
From plane
Where Eventually_within_30 inside(X_pos, P);
```

In Sistla et al. (1997), an algorithm to evaluate queries in the MOST model is presented. These queries are given in the form of conjunctive formulae. The proposed method of evaluation may support continuous spatio-temporal queries.

Update Policy and Uncertainty Management

The uncertainty aspect is very important in the field of moving object databases. Indeed the current location of moving objects is known only with limited accuracy. This is due to the continuous nature and the unpredictable behavior of the object movement, to the imprecision of positioning systems and to the network delays. We can consider the uncertainty notion at different levels. First, concerning the update of moving objects' current locations, an important issue is: "How often and when is updating the current location of a moving object needed in

order to have an acceptable uncertainty without affecting the system performance?" Second, it is necessary to take into account the uncertainty notion when querying the moving object location or trajectory. In the following, we present work that deals with these two aspects.

Moving Object Location Update

The first proposed approaches were based on periodic updates of moving object locations (e.g., every 2 km or every 2 minutes). These traditional approaches were used in commercial transportation (Wolfson, 2003). Their advantage is that they enable knowledge about the error bound in the query answers since the uncertainty is known in advance and it corresponds to the distance between two updates. However, if more accuracy on location is required, many more updates are needed. Thus, other update policies were proposed (Wolfson et al., 1999b). In this work the authors extend the MOST model by proposing an update policy based on deviation and uncertainty. The basic idea is to update the position of a moving object only when the distance between its current location and the database location exceeds a certain threshold. This distance is called deviation. It is computed by the moving object itself (e.g., equipped by GPS) or by sensors network. The choice of the suited uncertainty threshold is a very important issue. Indeed, the purpose is to find the compromise between the uncertainty cost and the update cost. So, three methods for determining the adequate uncertainty threshold were studied. The first one considers a fixed uncertainty threshold during the entire moving object trip. The second one changes the uncertainty threshold for each update in order to minimize the update cost, the deviation cost,

and the uncertainty cost. The third one deals with the disconnection detection. More details on these approaches are given in Wolfson et al. (1999a). Another recent approach was proposed (Wolfson, 2003). It is also based on deviation but it considers unknown itinerary contrary to hypotheses of Wolfson et al. However, the moving object is supposed to move on the same street between two updates.

Querying with Uncertainty

It is very important to deal with uncertainty in the data model, in the query language and execution model. In Wolfson et al. (1999a), the answers to queries involving current or future location of moving objects imply an uncertainty area (e.g., a circular area around the computed position having the uncertainty threshold as radius). In the case of DBMS representing the trajectory of a moving object, the problem of uncertainty management is more complex. Pfoser and Jensen (1999) have focused on the uncertainty due to sampling and to the imprecision of the GPS system. They propose a method that enables the quantification of uncertainty in moving object data modeling. Their approach enables limiting the uncertainty on the past movement of the moving object, but the error becomes higher when we are close to the present moment. Another relevant work considering the uncertainty on the trajectories of moving objects is Trajcevski, Wolfson, Zhang, and Chamberlain (2002). The authors propose to model the moving object trajectory as a cylindrical volume in 3D space.

OPEN PROBLEMS

In this section, we identify some open problems that seem to us to be the most relevant, and new challenges related to query processing in mobile environments.

New Scenarios

We have already classified the different query types related to mobility. Next, a very important question has to be asked about this subject: Do the query types studied meet all the requirements of the location-based applications? To our knowledge, the scenarios proposed so far do not take into account the ‘criticality’ aspect required by certain location-based applications. Assume, for instance, that a captain is looking for “the rescue helicopters able to reach his boat within 10 minutes.” Suppose also that this captain requires an answer in two minutes, another decision being taken otherwise. This query can be considered as a spatio-temporal query with real-time constraints. Indeed, we would like in that case for the query processing and result transmission to respect a particular deadline (two minutes). We can also imagine spatial constraints. A driver, for instance, could require obtaining an answer to an LDQ before reaching the next crossing or not farther than three miles from his current location. Thus, temporal deadlines or spatial constraints could both be specified in an explicit manner for different query types. In other cases, we could draw implicit constraints after the semantic analysis of a query. For example, let us consider spatio-temporal queries involving the future position of a moving object: select the ambulances which will be within 10 minutes of the hospital. In this case, we can consider the 10-minute delay as a deadline. Indeed, we can wonder what is the utility that will have the answer after that deadline?

We believe that new scenarios could open opportunities for new research that would propose methods to still open issues of processing mobile queries including real-time constraints. One of the interesting ways to do this would be to find a compromise between respecting the constraints and having a sufficient accuracy of

the information returned by the system. A user would sometimes prefer to have a result, even partial, that respects the deadline, rather than a complete result given after the specified deadline. In the example of the captain, he can accept a partial result (not all helicopters and not necessarily the closest) in order to have the answer in the two-minute delay. To our knowledge, such problems have not yet been addressed.

Unification of Approaches

There are various possibilities and ways of use, and the application domains are very wide. In fact, some applications provide location-dependent services to customers (e.g., local yellow pages). Others are tracking oriented (fleet management, transportation, emergency) or navigation-assistance oriented (e.g., inform a subscriber how to go from point A to point B). These applications have different requirements and different constraints. Consequently, they have been developed in several directions in order to fit differing location-based application requirements. Currently the trend is to develop tools and techniques dedicated to each application field. Regarding querying in mobile environments, the research was often led with the purpose of meeting the requirements and handling the constraints of particular types of queries. This has unfortunately led to a spread of energy and duplicate work. We believe that it is important, today, to generalize the approaches and to develop them in a unified framework in order to fit the common requirements of different types of applications. Hence, it could be easier to implement the approaches and even to standardize them. This effort has to be made at several levels (location management, data modeling, querying). We think that a relevant research direction would be to thoroughly study the constraints associated to each type of query,

with the aim of highlighting the common ones (e.g., the location, the movement, the temporal dimension). Such work would enable a common basis to all query types to be highlighted.

Standard Query Languages

For several types of queries studied in this chapter, we noted that querying involves specifying new operators. For spatio-temporal queries, in each proposed data model, new operators have been defined. Also for LDQ, we have seen that new operators taking into account proximity and orientation of the mobile client have been introduced. We think that it becomes necessary to enhance the “expressiveness” of the queries by standardizing these various extensions. Indeed, we must study formally the different proposals to integrate them in standard language intended for users (like SQL or OQL).

Mobile Queries Scheduling

In mobile environments, the determination of the queries scheduling, at the server side, requires some special considerations. Indeed, the answer to a query, submitted by a mobile client, could be sent back to him/her at a location that he/she would have already left. This problem requires, at best, re-routing the results to the new location of the client. This could be sufficient for NLRQ or LAQ, but not at all for LDQ. In fact, by changing location, the mobile client could receive invalid results, because answers to an LDQ are dependant on the client location. Hence, he/she would either reject the results or ask the server to process the query again. This could affect the system performance. Thus, we believe that an interesting direction to investigate would be to find a query processing order, taking into account the types of queries, their constraints, or the degree of mobility of the

client submitting them (high or low mobility). We think that future work could contribute to addressing this issue. Indeed, we can envisage the development of optimal scheduling algorithms of mobile queries, taking into account new constraints of mobility and different requirements of every type of query. This issue was also discussed in Lee et al. (2002).

Unpredictable Object Movement Modeling

We have outlined the fact that spatio-temporal queries may involve the anticipation of the future movement of objects. To process these queries, many authors assume that the movement of objects has certain regularity. In fact, the motion vector is determined by the positioning system. The position or the trajectories of mobile objects are often estimated by interpolations or extrapolations. Generally, these approaches are addressing applications that are moving object tracking oriented and navigation-assistance oriented. However, other applications that are information service oriented may require location prediction capability. Indeed, the integration of GPS modules in mobile phones and PDAs is increasingly spreading. Thus, new location-dependent commercial services may involve future positions of users having unpredictable movement. Indeed, the movement of a pedestrian using a mobile phone does not possess the same properties of regularity as other moving objects such as vehicles or planes.

CONCLUSION

In this chapter, we first proposed a classification of mobile queries into two categories: queries issued from mobile terminals and querying data related to fixed objects, and queries issued from mobile or fixed terminals and querying moving object databases.

These queries generated multiple problems, which created new fields of work. We focused on the research areas that were developed with the aim of resolving the problems related to query processing in mobile environments. We have chosen to present separately the proposed approaches within each of the two basic categories of query. For the first category we described the problems of location management. We also dedicated an important part to describing work related to location-dependent query processing. Concerning the second category of queries, we presented the most relevant work related to moving object databases modeling, especially their positions and their trajectories. We briefly described the work done with the aim of extending query languages in order to handle the spatial dimension and the temporal dimension. Finally, we were interested with the proposals made to face the problems of uncertainty management. Although progress is being made, there are still open issues which are challenging. In the last section, we presented some new challenges facing researchers before the use of complex mobile queries becomes an industrial and commercial reality.

REFERENCES

- Birman, K. P., Hayden, M., Özkasap, O., Xiao, Z., Budiu, M., & Minsky, Y. (1999). Bimodal multicast. *ACM Transactions on Computer Systems*, 17(2), 41-88.
- Chen, J., Dewitt, D. J., Tian, F., & Wang, Y. (2000, May). NiagaraCQ: A scalable continuous query system for Internet databases. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Dallas, TX (pp. 379-390).
- Erwig, M., Güting, R. H., Schneider, M., & Vazirgiannis, M. (1999). Spatio-temporal data

- types: An approach to modeling and querying moving objects in databases. *GeoInformatica*, 3(3), 269-296.
- ESRI. (2000). *About GIS: How GIS works*. Retrieved August 8, 2000, from <http://www.esri.com/library/gis/abtgis/giswrk.html>
- Forlizzi, F., Güting, R.H., Nardelli, E., & Schneider, M. (2000, May). A data model and data structures for moving objects databases. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Dallas, TX (pp. 319-330).
- Gök, H. G., & Ulusoy, O. (2000). Transmission of continuous query results in mobile computing systems. *Information Science*, 125(1-4), 37-63.
- Grumbach, S., Rigaux, P., & Segoun, L. (2000, September). Manipulating interpolated data is easier than you thought. *Proceedings of the International Conference on Very Large Data Bases*, Cairo, Egypt (pp. 156-165).
- Güting, R. H. (1989, August). Gral: An extensible relational database system for geometric applications. *Proceedings of the International Conference on Very Large Data Bases (VLDB)*, Amsterdam, The Netherlands (pp. 33-44).
- Güting, R. H., Bohlen, M. H., Erwig, M., Jensen, C. S., Lorentzos N. A., Schneider M., et al. (2000). A foundation for representing and querying moving objects. *ACM Transactions on Database Systems*, 25(1), 1-42.
- Holliday, J., Agrawal, D., & Abbadi, A. E. (2002). Disconnection modes for mobile databases. *Wireless Networks*, 8(4), 391-402.
- Kanellakis, P. C., Kuper, G. M., & Revesz, P. Z. (1990, April). Constraint query languages. *Proceedings of the Symposium on Principles of Database Systems*, Nashville, TN (pp. 299-313).
- Kanellakis, P. C., Kuper, G. M., & Revesz, P. Z. (1995). Constraint query languages. *Journal of Computer and System Science*, 51(1), 26-52.
- Leonhardi, A., & Kubach, U. (1999, October). An architecture for a universal distributed location service. *Proceedings of the European Wireless '99 Conference*, Munich, Germany (pp. 351-355).
- Lee, D. L., Lee, W. C., Xu, J., & Zheng, B. (2002). Data management in location-dependent information services. *IEEE Pervasive Computing*, 1(3), 65-72.
- Marsit, N., Hameurlain, A., Mammeri, Z., & Morvan, F. (2005, February). Query processing in mobile environments: A survey and open problems. *Proceedings of the International Conference on Distributed Frameworks for Multimedia Applications*, Besançon, France (pp. 150-157).
- Mokhtar, H., & Su, J. (2004, January). Universal trajectory queries for moving object databases. *Proceedings of the International Conference on Mobile Data Management*, Berkeley, CA (pp. 133-145).
- Mokhtar, H., Su, J., & Ibarra, O. H. (2002, June). On moving object queries. *Proceedings of the Symposium on Principles of Database Systems*, Madison, WI (pp. 188-198).
- Pfoser, D., & Jensen, C. S. (1999, July). Capturing the uncertainty of moving-object representations. *Proceedings of the International Symposium Advances in Spatial Databases*, Hong Kong, China (pp. 111-132).
- Pitoura, E., & Samaras, G. (2001). Locating objects in mobile computing. *IEEE Transactions on Knowledge and Data Engineering*, 13(4), 571-592.
- Serrano-Alvarado, P., Roncancio, C., & Adiba, M. E. (2001, September). Analyzing mobile

transaction supports for DBMS. *Proceedings of the International Workshop on Database and Expert Systems Applications (DEXA)*, Munich, Germany (pp. 595-600).

Seydim, A. Y., Dunham, M. H., & Kumar, V. (2001a, May). Location-dependent query processing. *Proceedings of the International Workshop on Data Engineering for Wireless and Mobile Access*, Santa Barbara, CA (pp. 47-53).

Seydim, A. Y., Dunham, M. H., & Kumar, V. (2001b, September). An architecture for location-dependent query processing. *Proceedings of the International Workshop on Database and Expert Systems Applications (DEXA)*, Munich, Germany (pp. 549-555).

SignalSoft. (2000). *SignalSoft Corporation—Wireless location services*. Retrieved December 2, 2000, from <http://www.signalsoftcorp.com>

Sistla, A. P., Wolfson, O., Chamberlain, S., & Dao, S. (1997, April 7-11). Modeling and querying moving objects. *Proceedings of the International Conference on Data Engineering*, Birmingham, UK (pp. 422-432).

Trajcevski, G., Wolfson, O., Zhang, F., & Chamberlain, S. (2002, March). The geometry of uncertainty in moving objects databases. *Proceedings of Advances in Database Technology—EDBT, International Conference on Extending Database Technology*, Prague, Czech Republic (pp. 233-250).

Wolfson, O. (2003, July). Accuracy and resource consumption in tracking and location prediction. *Proceedings of the International Symposium on Advances in Spatial and Temporal Databases*, Santorini Island, Greece (pp. 325-343).

Wolfson, O., Jiang, L., Sistla, A. P., Chamberlain, S., Rishe, N., & Minglin, D. (1999). Databases for tracking mobile units in real time. *Proceedings of Database Theory—ICDT, International Conference*, Jerusalem, Israel (pp. 169-186).

Wolfson, O., Sistla, A.P., Xu, B., Zhou, J., & Chamberlain, S. (1999, June). DOMINO: Databases fOr MovINg Objects tracking. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Philadelphia (pp. 547-554).

Section IV

Network

Chapter XX

A Neural Network–Based Mobile Architecture for Mobile Agents

Anand Kuppuswami

University of Western Sydney, Australia

ABSTRACT

Wide area network (WAN) offers advantages like providing myriad services available on globally diversified computers with reasonably simple process. The ability to dynamically create networks offers the processing powers of various processors at our command. With the advent of protocols like SOAP and Web services, the consumption of services are more organized. In spite of various advances in communication techniques, the consumption of services through mobile gadgets is still only at the research level. The major impedances in implementing such systems on a mobile network are the latency factor, abrupt disconnection in service, lower bandwidth and minimal processing power. The mobile agent's paradigm proves to be an effective solution to various issues raised. It has received serious attention in the last decade and several systems based on this paradigm have been proposed and built. All such systems have been designed for a static network, where the service providers and the requestors are connected to the server on a permanent basis. This chapter presents a new framework of managing the mobile environment and the participating nodes with active intelligent migration. The functioning of the mobile agents in such a scenario is also presented.

INTRODUCTION

Mobile devices foster a new set of applications that are receiving enormous support in the global electronic community. This is primarily because of the ability of mobile devices to connect and reconnect with each other dynamically. The ability to create a network of devices “on the fly” can be adjudged as a major advan-

tage of mobile gadgets as compared with the previous computer networks. The ability of being able to communicate and, more importantly, process information irrespective of time and place is a very promising feature offered by mobile technologies, as also ratified by Gray, Kotz, Nog, Rus, and Cybenko (1996). With the advent of the Internet and, especially, in communication, almost all actions related to an

application have been reduced to a few mouse clicks. As is obvious in our day-to-day usage, the Internet is a vast repository of information and services. Even though both Internet and related mobile technologies are very promising, their usage by business is still full of hurdles. For example, accessing a simple HTML page in a mobile gadget requires various configurations in them. This can be attributed to low latency, physical obstruction, and network connectivity of the mobile gadgets. The mobile gadgets have low memory, processing capabilities, and are prone to sudden failure (Jipping, 2002). These factors unveil the problem of client server architecture (Gray et al., 2002b). Often, having a well-thought-out networking architecture is crucial to successful usage of this technology. For example, the huge amount of data that is required to be transported due to various multimedia applications to the client's site for further processing is not possible without a good architecture. This created the need for remote working where code could migrate to a different machine, execute at the new machine, and return with the result. The whole concept of working remotely started with the message passing and remote procedure calls (RPCs) provided by Java (Birrell & Nelson, 1984). But their application was limited, as the client could use only those services provided by the server. If the requested service is not present, then the client has to make intermediate or "Bridge" function class in order to get to the final result. This process results in wastage of resources and bandwidth. As an alternative approach, small subprograms could be written and passed on to the service provider to execute locally. Network Command Language (NCL) (Meandzija, 1986), remote evaluation (REV) (Stamos & Gifford, 1990), and SUPRA-RPC (Stoyenko, 1994) employ this idea in their architecture. All these architectures had one major concern: the code, once migrated to a different machine, cannot remigrate at the end of execu-

tion. The architecture also lacked in active coordination between different programs. The absence of inter-communication protocols leads to poor performance of the system.

The novel approach of transportable programs (Gray, 1995; Cybenko, Gray, Wu, & Khrabrov, 1994) offers a promising solution for various issues raised. Transportable agents or mobile agents, as they are called now, are autonomous programs that can migrate from one machine to another machine in the network. By migrating to the machine having the resource, the agents have the advantage of working on site where the resource is present and also use the processor's power. This eliminates all the middleware that is required for transporting the data to the client's site. The mobile agent's paradigm provides an effective solution to the problem of low latency, poor interface, and bad network conditions (Gray et al., 1996). The middleware and the communication control mechanism form the major workload in a client-server architecture; by eliminating them, we can build a better working environment and increase the efficiency of the system. The code and state of code is migrated to another machine for resuming its execution. This also eliminates the interface required for service access. The fact that there is no need for permanent connection makes it very suitable to the mobile environment. The ad hoc client-server model is overridden by the peer-peer model which matures into grid computing, where the machines can act as client or server depending on the environment. The programmer is swayed away from traditional multi-tier architecture to grid computing (Lauvset, 2001).

The majority of the mobile agents present in the literature is designed for a static network architecture. Baring a few mobile agents (Cabri, Leonardi, & Zambonelli, 2002; Kendall, Krishna, Pathak, & Suresh, 1998), intelligence is not embedded into them. We present a new set of mobile agents which work in a volatile mobile

environment. Embedded with intelligence, agents can act autonomously, collaborate with other agents, and reduce the human intervention. The agents are configured to be adaptable, learn from their experience, and mature into experienced agents.

REVIEW OF MOBILE AGENTS AND THEIR LIMITATIONS

A new software architecture for building distributed applications using the mobile agent paradigm, which also supports cross-referencing between agents using Telescript (Dömel, 1996), has been proposed by General Magic. The Telescript agents can begin their execution in any machine, migrate to a different machine, and continue their execution, with the only prerequisite being that all hosts are Telescript enabled. Telescript is an object-oriented language that is a collection of hierarchically organised classes. The Agent class, which is inherited from the Process class, has a method “go”, which performs the migration process. When an agent wants to migrate to another machine, it issues the command “go”. The agent continues its execution from the command following the “go” command. A server at each site authenticates (Tardo & Valente, 1996) the agents and executes them. Requirements like special hardware and support for single language limits the global implementation.

Tacoma (Johansen, Renesse, & Schneider, 1996)—developed at the University of Tromsø and the University of Cornell—are agents written in TCL/HORUS, which is a version of TCL where the HORUS is used for group communication and fault tolerance. They have rear guards which restart lost agents and have features like electronic cash for services. They have protection mechanisms which ensure that there are no runaway agents. Broker services for scheduling and directory services are also

implemented. In spite of several advantages, the programmer has to explicitly capture the state information and does not support interruption of the execution of program, which limits its usability.

Agent TCL—D’Agents as it is now called (Gray, Cybenko, Kotz, Peterson, & Rus 2002a)—is a mobile agent system which attempts to strike a balance between the Tacoma and Telescript systems. It uses the flexible scripting language, TCL (Isaacson, 2001) and also provides support for other languages. Various migration and communication primitives help the programmer to code without worrying about the low-level migration protocol and also gives enough granularities for building basic blocks. For security purposes, it uses the SafeTCL (Levy, Ousterhout, & Welch, 1997) protocol. In the Agent TCL (Gray, 1998; Kotz et al., 1999) environment, the agents can migrate from one machine to another, create child agents for performing sub-tasks, access resources across networks, and communicate with other agents residing in local or remote machines. TIAS (Transportable Intelligent Agent System) (Harker, 1995) is an improved version of D’Agents, and offers several patches over existing systems.

A new series of itinerant agents (Chess, Grosz, Harison, Levine, & Parris, 1995) has been proposed which offers a secure environment for remote applications in public networks. Itinerant agents are programs which are dispatched from a source computer to roam among a set of networked computers until they accomplish the required task. The need for knowledge representation and verification is also discussed in their architecture. The Messenger-based operating system (Marzo, Murhimanya, & Harms, 1994) and some work on creating new mobile agents from existing systems (Brazier, Overeinder, Steen, & Wijngaards, 2002) have led to creating mutants like TOMAS (Transaction-Oriented Multi-

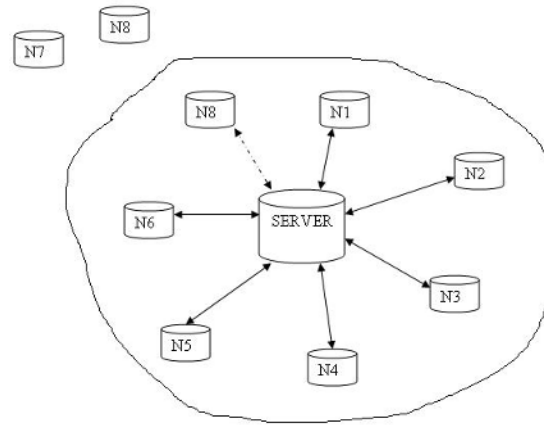
Agent System) developed by Busetta and Ramamohanarao (1998).

With the working knowledge of various systems, and taking into account their inadequacies, we propose to build a new framework that offers a promising edge over the existing systems. The framework's working includes various modules, like knowledge building, pattern recognition, and intelligent migration. The architectures based on mobile agents' paradigm vary with one another in several aspects. The process of comparison requires concrete understanding of the terms like desired degree of configurability, scalability, and customizability (Fuggetta, Picco, & Vigna, 1998). The lack of such standardization subdues any work on comparison.

MANAGING MOBILE ENVIRONMENT

There are several issues that have to be addressed before a mobile environment can be implemented. For example, current location of the mobile agent needs to be clearly identifiable. This is required in situations where some parameters need to communicate to the agent. A service that is published in the UDDI directories can be easily availed with the help of Web services (Yang, Hsieh, Lan, & Chung, 2005). In certain cases where the services are not published in the required format, there needs to be some brokering before the services can be consumed (Cybenko & Jiang, 1999). These could be achieved by those brokering, which works as the middleware. One of the most important aspects that needs to be considered is the current working environment. If the nodes in the network are volatile, then the process of migration gets intricate. As the nodes are volatile, the services also offered variegates. Presence of certain nodes may

Figure 1. A sample mobile network



result in the system offering better service, and their absence may force one to look for the second best choice. In the current scenario, there needs to be an active broker, who advises the agent on its migration. As the service offered cannot be compared on a binary basis, there need to be “human aspects” embedded into the broker service.

We propose to utilize the agent tracking feature of architectures like Agent TCL and Telescript, where a central system keeps track of the active agents and their locations. A central server keeps tracks of the service providers. The use of knowledge base is also incorporated into our system. Depending on current requirement, intelligence may or may not be required. In situations where simple binary logics will not suffice, we propose to use advanced models like neural networks; probabilistic methods offer a promising solution.

Introduction to Mobile Environment

Figure 1 presents a sample mobile network where several nodes offering various services are collaborating with each other and the central server. The services are hexa-coded using

keywords, which are then registered with the server. In the static scenario all the nodes are permanently connected to the server, whereas in the mobile environment, the nodes will be connected for a shorter time span. The number of nodes connected to the server will also fluctuate. To explain the working of a mobile environment, let us assume that we have eight nodes, of which six are connected to the server. The agents are in active collaboration mode, which could be productive or counter-productive depending on the process.

The static architecture is more like a simple market environment, where we have all the shops and we visit them to consume the services offered. In some cases, the set of current service providers do not offer the solution we need, and we have to settle with the level of services offered. A more generalised architecture would make the service providers mobile. The service providers “walk in and walk out” of the market anytime, at their will. This results in a market that would be highly volatile. There could be a set of nodes that remain connected and offer their services on a more permanent basis, but other nodes could be volatile. The presence or absence of the volatile nodes can have a vast impact on the functioning of the system.

Migration of Nodes

Each node has a header section called the Service Description Header (SDH), as shown in Figure 2, which describes the node in detail. The descriptions include the node name, the speed of connection, the quality of service factor, services offered, proximity index (PI), the operating system, and the language of coding. Whenever the node wants to enter the working environment, it sends a request to the server. The request would contain information like the START_TIME, END_TIME (if appli-

Figure 2. A sample service description header

```

<ROOT>
<START_TIME>
12:15
</START_TIME>
<END_TIME>
NA
</END_TIME>
<SERVICE>
<NAME>
IMAGE
</NAME>
<PI>
14
</PI>
</SERVICE>
<SERVICE>
<NAME>
PROCESSING
</NAME>
<PI>
10
</PI>
</SERVICE>
<SERVICE>
<NAME>
NEURAL
</NAME>
<PI>
11
</PI>
</SERVICE>
<SERVICE>
<NAME>
NETWORK
</NAME>
<PI>
8
</PI>
</SERVICE>
</ROOT>

```

cable), and the broad category of services. The PI would be a number ranging from 0 to 15, signifying how close the service is to the respective keyword. This would be in XML format for compatibility in heterogeneous environment.

Based on this request, the server assigns a unique ID to this node. In some cases this service may be redundant, and in that case server places the node in the queue. Once the node has been assigned an ID, the node prepares the detailed service description. The description will be in XML format and would have the structure as shown in Figure 3.

Similarly, whenever the node wants to exit the operating environment, it sends a termination request to the server. In case of termination, if any agent is being executed in that node, then the termination is postponed till the agent finishes the execution. Based on the START_TIME and the END_TIME, a latency factor and a quality of service factor are assigned to each node.

While assigning a new ID to the node entering the operating environment, the server checks for the existence of previous ID. If the node already had an ID, a new ID is not created and the old ID is retained. While assigning the new ID, the server checks the history for latency

Figure 3. A sample service description in XML format

```

<ROOT>
  <ID>
    ABC123
  </ID>
  <OS>
    UNIX
  </OS>
  <LANGUAGE>
    C++
  </LANGUAGE>
  <SERVICE>
    <NAME>
      IMAGE
    </NAME>
    <PI>
      14
    </PI>
  </SERVICE>
  <SERVICE>
    <NAME>
      PROCESSING
    </NAME>
    <PI>
      10
    </PI>
  </SERVICE>
  <SERVICE>
    <NAME>
      NEURAL
    </NAME>
    <PI>
      0
    </PI>
  </SERVICE>
  <SERVICE>
    <NAME>
      NETWORK
    </NAME>
    <PI>
      8
    </PI>
  </SERVICE>
  <SERVICE>
    <NAME>
      HISTOGRAM
    </NAME>
    <PI>
      12
    </PI>
  </SERVICE>
  <SERVICE>
    <NAME>
      PATTERN
    </NAME>
    <PI>
      0
    </PI>
  </SERVICE>
  <SERVICE>
    <NAME>
      HIERARCHICAL
    </NAME>
    <PI>
      14
    </PI>
  </SERVICE>
  <SERVICE>
    <NAME>
      ADAPTIVE
    </NAME>
    <PI>
      15
    </PI>
  </SERVICE>
  <ROOT>

```

factor and the quality of service factor. If the values are lower than the values existing in the server's database, then the lower values are substituted; otherwise, the old values are retained. This process ensures that the quality or the latency factor is never hyped. Once the process of registration is complete, the node becomes an active member.

Challenges of Node Migration

In each instance, prior to the migration of agents between different nodes, a few factors must be considered in detail. Whenever the code migrates, it needs to take with it the working variables, so that it can continue the execution in the new environment. This brings up the question: How much data needs to accompany the code? Let us consider a simple problem where a mobile agent is required to calculate the amount of tax that needs to be deducted from a salary. In this case, a numerical value is all that needs to accompany the code in the process of migration. This numerical value is enough for the successful execution

of the code in the new environment. In this case the service provider has more information than the service requester. In a homogeneous environment, this process is further simplified. Now let us assume a situation in which the image format needs to be changed. For example, we have an image in BMP format and would need to convert it into JPEG format. Here, the mobile code has to transport the whole image to the site of service provider. This suggests that the migration of code would consume more resources than the service provider passing on the service. So we have to calculate the amount of data and the bandwidth of operation, and depending on those values, we decide on the migration.

A NEURAL NETWORK-BASED AGENT FOR INTELLIGENT MIGRATION

In the mobile environment, the nodes are volatile and hence the process of migration cannot be hard-coded. We need a run-time tool that can decide on the address of the future node. Each node varies on the service provided and has a varying level of PI associated with it. This perplexes the migration process. The migration also needs "shrewd attitude" when analysing the service factors. With these additional responsibilities thrust upon the mobile agent, the need for an advanced intelligent assistant arises. We propose an agent called the *broker agent* which performs the decision on migration of agents and acts as the "human intervention."

Request of Service and Broker Agent

When the mobile agent execution stack gets to the "go" command, the mobile agent creates a broker agent (BA). The purpose of the broker

Table 1. A sample keyword list

Keyword	Comments
Image	Deals with image operations
Processing	Process images
Histogram	Performs histogram of the image
Equalization	Deals with histogram

agent is to communicate with the server and identify the possible nodes which can offer the required service. The BA will communicate the agent's requirement to the server. This would result in the server identifying more than one node that can offer the required service. The BA must further drill down on the search result and track down the best node.

The list of the keywords and the service description of each node will result in varying system usability. In such a scenario a simple binary logic would not suffice. For example, if the service provider says "IMAGE EQUALIZATION" in the list of services provided and the agent is looking for "HISTOGRAM EQUALIZATION," then the simple binary logic will turn down the listed service. But in the image processing arena, both terms are used synonymously.

Ontology of Keywords and Network Architecture

As a first step, a list of keywords and their corresponding binary codes are prepared. This is presented to the neural network as inputs. The structure of the network is such that it can accommodate future unpredicted inputs. We would need multiple networks for performing multiple tasks. As each of these services does not have a maximum limit on number of keywords, we need a system which can handle this requirement. We suggest a network architecture of 16 input nodes which can accept an input between 0 to 15 for each node. This leads to an

infinitely large number of a combinations of input values. The actual structure of the input is explained in the following section. Depending on the complexity of the service required, a *keyword list* (KL) is generated. A sample KL is shown in Table 1. The comments in the second column only aid the expert in his process of building the KL. This is not used in any part of the automated processing.

For example, let us assume that we need to perform histogram equalization. Then the list of keywords could be histogram, equalization, image, processing, and intensity. Each of them will become the input to the network. This suggests that we need five input nodes. This architecture will successfully work in an environment where the keywords are not emerging and fail otherwise.

To tackle this issue we propose to create agents having architecture which can cater for future keywords. The input to each node would be a value between 0 and 15. We also have decimal inputs which can help in getting accurate solutions. The sample input to the agent for training would be as in Table 2, where we have only shown whole numbers as the input, but decimal values are also accepted.

The input to the agent would depend on the proximity index for each service. For example, if the service provider has the keywords "image" and "processing", and their corresponding PI to be 15 and 8, this suggest that the input to the agent would be "F800 0000 0000 0000." We can map the service provider's PI to the agent input as shown in Table 2

Table 2. A sample input to the broker agent

Service Provider Keyword PI	Input to Neural Network
Image 15, Processing 8	F800 0000 0000 0000
Histogram 15, Processing 8	08F0 0000 0000 0000
Histogram 15, Equalization 10	00FA 0000 0000 0000
Image 15, Processing 15, Equalization 8	FF08 0000 0000 0000
Image 15, Processing 15, Histogram 15, Equalization 15	FFFF 0000 0000 0000

Table 3. Process of mapping

Keyword	Action
Photo	Mapped to Image
Function	Mapped to Processing
Normalization	New entry created
PCA	New entry created
Picture	Mapped to Image

Mapping Keywords

The process of mapping is complex and needs expertise at various stages. When mapping the PI for the known keywords, the process is relatively simple. The mapping complicates when the agent encounters unknown keywords. An expert's knowledge would then be required to map the new keyword to either an existing one or to add the keyword to the KL. For example, if the service provider has "photo" as the keyword, then we can see that it is very closely related to the keyword "image". In this case, there would be no need to add a new keyword. Now let us suppose that the service provider has a keyword "PCA". Since it is not directly related to any of the existing keywords, there would be a need to alter the KL. A sample scenario is shown in Table 3.

Depending on the agent's service required, the KL may be a huge data repository in itself. This process of mapping cannot be automated at this stage of development.

Training of Network

Initially, when training the network only the five bits are used for training. The set of output values for each set of input has to be decided by an expert. This is a supervised training, which requires the knowledge and experience of a system expert. The output of the system can vary from -1 to +1, signifying how well the service provided suits the requirement. For example if the service provider has the keywords "image" and "histogram", then it has a better chance of catering to the needs when compared to a service which only has "image" as its service keyword. Depending on the PI for each keyword, the input value to the agent varies. A sample set of input and corresponding output can be formulated as shown in Table 4.

This input and output set is employed to train the network using the back propagation algorithm. After the successful completion of training, the neural network will be able to predict the outputs for unknown inputs. The node which

Table 4. A sample training set to the broker agent

Input to Neural Network	Output
F800 0000 0000 0000	+0.5
08F0 0000 0000 0000	+0.8
00FA 0000 0000 0000	+1.0
FF08 0000 0000 0000	+1.0
FFFF 0000 0000 0000	+1.0
F800 0000 0000 0000	-1.0

evaluates to the highest value is selected as the service provider.

Process of Agent Learning and Maturing

After the initial phase of training, the agent is released to the network of nodes. The agent starts evaluating each node's service. The BA communicates with each node, gets its SDH, and extracts the keywords and their PIs. Depending on the keyword patterns, each node is evaluated for its service against the agent's requirements. The node having the highest match or the least conflict is selected as the service provider. In this process, if new keywords are encountered, then the agent would be unable to evaluate the performance of that node. In this case, new keywords need to be first mapped before the agent can process them. The keywords need to be mapped before the BA can proceed further. If there is no change in the KL, then without further processing the values are presented to the agent. If there is a change in the KL, the network needs to be retrained using the new sets of input. This is the process where the agent starts learning and maturing. After a series of training, the agent would have matured enough, and would not need further training and can be fully used. This process can be compared to a new staff member in an office who needs a few sessions of training before he can act independently.

IMPLEMENTATION OF INTELLIGENT MOBILE AGENTS

The system is being implemented in the .NET environment using C#, IIS as the Web server, XML, and SOAP protocol for communication. The project has been divided into four phases. In Phase I, the service request module is developed. In Phase II, the mobile environment is developed. In Phase III, the intelligent agents are developed, and in Phase IV, all the modules are integrated and fully implemented. Here we are trying to develop an application where a mobile client captures the image of a person and wants to find more information about him. This would be the case where a security person on his regular rounds finds a suspect but is not sure. In this situation we cannot expect a permanent connectivity with the server. So the ideal situation would be to capture the image and send it to the ground station for further processing. In a more generalised scenario, we can say that any person can capture the image and pass it on to the government agency for further scrutiny. Here there would be a series of activities that need to be performed before the information can be retrieved. The list of activities could range from capturing of image, communicating with the ground equipment, creating the broker agent, selecting the node for service, performing the migration, and processing the image.

All of the activities would have a list of tasks that needs to be completed. For example, image processing would need to perform tasks like image clarity improvement, image format conversion, image clipping (to get rid of background information), image histogram equalization, and image PCA. Each of these tasks would be performed by individual agents working in their environment.

In this chapter we describe the task of performing histogram equalization. Initially the agent is trained as proposed. Then the node communicates with all the nodes in the network suggested by the server. The list of keywords are analysed and the KL is updated. Depending on the KL change, the agent is retrained; this process continues until there is no further change in the KL. This would suggest that the agent has matured enough to perform the operation autonomously. Then the agent starts evaluating the nodes for its service index. The node which evaluates the highest value is selected for the service performance. Phase I has been developed, and Phase II is in the development stage.

CONCLUSION AND FUTURE WORK

We have presented a new framework of mobile agents functioning in a volatile mobile environment. The mobile environment is the generalised situation in the real world where the service providers oscillate between providing and revoking their service. We have presented a managing technique which can perform various tasks relating to managing such an environment. The agents have intelligence embedded into them in the form of neural networks, which acts in a "human way." The agents are initially trained and then let to mature by iterative retraining. This system is presently in the development stage and has several aspects that need

further experimenting. Features such as the process of calculating the latency factors for each node are still in the research stages and need experimental data for analysis.

REFERENCES

- Birrell, A. D., & Nelson, B. J. (1984). Implementing remote procedure calls. *ACM Transactions on Computer Systems*, 2(1), 39-59.
- Brazier, M. I., Overeinder, B. J., Steen, M., & Wijngaards, N. J. E. (2002). Agent factory: Generative migration of mobile agents in heterogeneous environments. *Proceedings of the 2002 ACM Symposium on Applied Computing (SAC 2002)* (pp. 101-106).
- Busetta, P., & Ramamohanarao, K. (1998). An architecture for mobile BDI agents. *Proceedings of the 1998 ACM Symposium on Applied Computing* (pp. 445-452).
- Cabri, G., Leonardi, L., & Zambonelli, F. (2002). Engineering mobile agent applications via context-dependent coordination. *IEEE Transactions on Software Engineering*, 28(11), 1039-1055.
- Chess, D., Grosz, B., Harison, C., Levine, D., & Parris, C. (1995). Itinerant agents for mobile computing. *IBM Research Report on Computer Science and Mathematics* (RC 20010).
- Cybenko, G., & Jiang, G. (1999, July 18). Matching conflicts: Functional validation of agents. *Proceedings of AAAI 99*, Orlando, FL.
- Cybenko, G., Gray, R. S., Wu, Y., & Khrabrov, A. (1994). *Information architecture and agents*.
- Dömel, P. (1996). Mobile Telescript agents and the Web. *Proceedings of the 41st IEEE International Computer Conference* (p. 52).

- Fuggetta, A., Picco, G. P., & Vigna, G. (1998). Understanding code mobility. *IEEE Transactions on Software Engineering*, 24(5).
- Gray, R. S. (1995, December). Agent TCL: A transportable agent system. *Proceedings of the CIKM Workshop on Intelligent Information Agents of the 4th International Conference on Information and Knowledge Management (CIKM 95)*, Baltimore, MD.
- Gray, R. S. (Ed.). (1998). *Agent TCL: A flexible and secure mobile-agent system*. Hanover, NH: Dartmouth College.
- Gray, R. S., Cybenko, G., Kotz, D., Peterson, R. A., & Rus, D. (2002a). D'Agents: Applications and performance of a mobile-agent system. *Software—Practice and Experience*, 32(6), 543-573.
- Gray, R. S., Kotz, D., Nog, S., Rus, D., & Cybenko, G. (Eds.). (1996). *Mobile agents for mobile computing*. Hanover, NH: Dartmouth College.
- Gray, R. S., Kotz, D., Peterson, R. A., Barton, J., Chac, D. A., Gerken, P. et al. (2002b). Mobile agent versus client/server performance: Scalability in an information-retrieval task. *Proceedings of the 5th International Conference on Mobile Agents* (pp. 229-243).
- Harker, K. (Ed.). (1995). *TIAS: A Transportable Intelligent Agent System*. Hanover, NH: Dartmouth College.
- Isaacson, P.C. (2001). Object-oriented programming in TCL/TK. *Journal of Computing Sciences in Colleges*, 17(1), 206-215.
- Jipping, M.J. (2002). *Symbian OS communications programming*. New York: John Wiley & Sons.
- Johansen, D., Renesse, R., & Schneider, F. B. (1996). Supporting broad Internet access to TACOMA. *Proceedings of the 7th ACM SIGOPS European Workshop* (pp. 55-58).
- Kendall, E. A., Krishna, P. V. M., Pathak, C. V., & Suresh, C. B. (1998). Patterns of intelligent and mobile agents. *Proceedings of the 2nd International Conference on Autonomous Agents* (pp. 92-99).
- Kotz, D., Gray, R., Nog, S., Rus, D., Chawla, S., & Cybenko, G. (1999). *Agent TCL: Targeting the needs of mobile computers* (pp. 513-523).
- Lauvset, K. J. (2001). Separating mobility from mobile agents. *Proceedings of the 8th Workshop on Hot Topics in Operating Systems* (p. 173).
- Levy, J. Y., Ousterhout, J. K., & Welch, B. B. (Eds.). (1997). *The Safe-TCL security model*. Sun Microsystems.
- Marzo, G. D., Murhimanya, M., & Harms, J.R. (1994). *Messenger-based operating systems*. Technical Report No 90, Cahier du CUD, University of Geneva, Switzerland.
- Meandzija, B. (1986). A formal method for composing a network command language. *IEEE Transactions on Software Engineering*, 12(8), 861-865.
- Stamos, J. W., & Gifford, D. K. (1990). Remote evaluation. *ACM Transactions on Programming Language Systems*, 12(4), 537-564.
- Stoyenko, A.D. (1994). SUPRA-RPC: SUBprogram PaRAmeters in remote procedure calls. *Software—Practice and Experience*, 24(1), 27-49.
- Tardo, J., & Valente, L. (1996). Mobile agent security and Telescript. *Proceedings of the 41st IEEE International Computer Conference* (p. 58).

Yang, S. J. H., Hsieh, J. S. F., Lan, B. C. W., & Chung, J.-Y. (2005). Composition and evaluation of trustworthy Web Services. *Proceedings of the IEEE International Workshop on Business Services Networks* (p. 5).

Chapter XXI

Load Balancing as a Key to Enable Different Services in a Cellular Network

Sipra Das(Bit)

Bengal Engineering and Science University, India

Sulata Mitra

Bengal Engineering and Science University, India

ABSTRACT

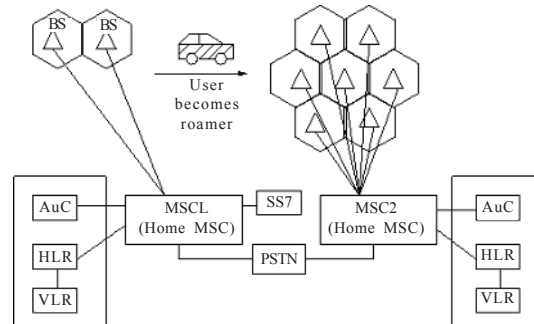
This chapter develops the concept of load balancing that plays a key role in providing various advanced application in cellular mobile environment. Load balancing means the efficient distribution of channels among cells in accordance with their requirements to minimise call blocking. As the channels for these services are scarce, load balancing has emerged as a primary issue in today's scenario. Two different prominent schemes of load balancing are elaborated. This chapter is aimed at the researchers and the policy makers making them aware of the different means of efficient load balancing as well as underscoring the problem areas that need further vigorous research.

INTRODUCTION

With the advancement of technology and behavioural change in work patterns/tasks, the world of computing has been experiencing increasing change in several directions. More particularly, the recent advances of cellular systems have rendered geographical distances insignificant. The architecture of a cellular

network, as shown in Figure 1, is generally conceived as a collection of geometric areas called cells. Each cell is serviced by a base station (BS) located at its centre. A number of BSs are again linked to a mobile switching centre (MSC) acting as a gateway of the cellular network to the existing wired network, such as PSTN, ISDN, and so forth. The wireless communication takes place only between

Figure 1. Basic cellular network



the BS and the mobile user (MU). Other communications—for example, between BS and MSC or between MSC and MSC—are wire lines (Lee, 1995; Rappaport, 1999; Boucher, 2001).

For the last decade or so, the number of mobile users has grown enormously, and as very limited frequency spectrum is allocated to this service, the efficient sharing of the spectrum among the users has become an important issue. In order to combat the problem arising out of limited allocation of frequency spectrum, the frequency channels are reused as much as possible to support a huge number of simultaneous communicating users. To increase the utilisation of limited frequency, further various schemes of channel allocation are developed (Tajima & Imamura, 1988; Del Re Fantacci, & Giambene, 1995; Das, Sen, & Jayaram, 1998; Das(Bit) & Mitra, 2000; Mitra & Das(Bit), 2000); two are presented here. In these schemes, available frequency channels are efficiently assigned and/or borrowed, which consequently reduces the number of blocked calls in each cell.

Load balancing means the efficient distribution of channels among cells as per their requirement to minimise the call blocking probability. This can be done either by assigning the frequency channels among the cells and/or by borrowing channels from cells having excess

channel. In the first scheme (Das(Bit) & Mitra, 2000) presented here, a load balancing technique is described maintaining a few small databases based on the history of traffic for a recent certain period. It is basically a combined approach, marrying a fixed and dynamic channel assignment in load balancing. Initially all the cells get some fixed number of channels. Additionally there is a central reserve pool managing the extra demand of channel (if any). The merit of this semi-dynamic-type channel assignment scheme is that message trafficking among different components of the network is much less with respect to the existing schemes, which are purely based on fixed assignment technique. As the message trafficking is reduced, more channels are made available, reducing call blocking probability.

The second scheme (Mitra & Das(Bit), 2000) is a combination of dynamic channel assignment and channel borrow technique. Channels are assigned dynamically among the cells under each MSC. In cases in which channel demand may increase, the cell that needs excess channels due to the occurrence of event can borrow a channel from one of its compact pattern (CP) cells by exchanging a few number of messages. If the probable lender cell is not available, the MSC can borrow channels from other MSCs that have excess channels.

BACKGROUND

Several schemes have been proposed to assign channels to the cells so that available spectrum is efficiently used. The existing assignment schemes may be broadly classified as fixed (Macdonald, 1979), dynamic (Del Re et al., 1995), and flexible (Tajima & Imamura, 1988; Das(Bit) & Mitra, 2000) schemes. Other than the channel assignment scheme, some schemes are also proposed, taking channel borrowing (Das et al., 1998; Mitra & Das(Bit), 2000) into consideration.

In the fixed assignment scheme (Macdonald, 1979), a set of frequencies is statically allocated to each cell. Although the fixed assignment schemes perform well under heavy traffic conditions, one major drawback is that if the number of calls exceeds the set of channels assigned for the cell, the excess calls are blocked.

The dynamic channel assignment techniques are used to increase the utilisation of available channels. In such systems, initially no channel is assigned to the cells in the system. But a cell can send a request for a free channel if necessary. The system keeps all the excess channels in a pool. As soon as a channel release occurs in a cell due to the termination of a call, channel rearrangement must be done.

In one such dynamic channel allocation scheme (Del Re et al., 1995), when a call arrives in a cell, the scheme starts to work either at a mobile switching central office or at a single base station, depending on whether the system is centralised or decentralised respectively. This scheme takes care of hand-off calls along with the new calls. In this scheme, the blocking probability for hand-off calls is lower than that for new call attempts, thereby providing better service quality. The dynamic channel allocation scheme is expected to cope better with traffic overloads to a certain extent, but on

high demands, the computational overheads for this scheme defeat the purpose of the scheme.

To reduce the computational overhead, flexible channel assignment schemes (Tajima & Imamura, 1988; Das(Bit) & Mitra, 2000) are proposed. The flexible channel assignment schemes combine the concept of both fixed and dynamic assignment schemes. There is a fixed set of channels for each cell, but channels are also allocated from a global pool in case of any shortage.

Although the objective of all the schemes is better utilisation of the available channels, causing reduced call blocking probability, most of the schemes do not consider the non-uniformity of channel demand. Here non-uniformity means the variation of channel demand from time to time in a cell and/or from area to area. For example, in the downtown area of a city, channel demand may be high during office hours. Hence one of the abovementioned schemes (Das(Bit) & Mitra, 2000), which follows temporal non-uniformity, is elaborated in a later section.

To combat the problem of temporal as well as spatial non-uniformity of channel, channel borrowing techniques (Das et al., 1998; Mitra & Das(Bit), 2000) are proposed. In one such scheme (Das et al., 1998), a hot cell in an inner ring can borrow channels only from a cell in the adjacent outer ring. This lending mechanism minimises the co-channel interference. This scheme transfers channels through a structured borrowing mechanism from the cold cells to the hot cells constituting the hot spot. But the scheme suffers from the overhead of huge message exchange between the mobile switching centre and base stations.

LOAD BALANCING SCHEMES

Among the abovementioned schemes, two load balancing schemes proposed by the authors are

elaborated on in the following subsections. The first scheme is named Scheme I and the second is named Scheme II.

Load Balancing Scheme I

This is a flexible channel assignment strategy (Das(Bit) & Mitra, 2000) where MSC maintains two databases, one to keep record of the extra channel demand and the other to store the call-block history at various cells under it. After each update period, the database that keeps the extra channel demand is self-updated with the help of the other database that maintains the call-block history. Along with these two databases, each MSC maintains some information at a central reserve pool. Assisted by this information, the MSC assigns channels from the pool to the individual cell to fulfil extra demand (if there is any).

The Database at MSC

Each MSC maintains two databases. One database named EXT_DEM keeps cell-wise extra demand of channels arising out of overload of traffic in different time slots. The database schema of the EXT_DEM is (cell_id, time_stamp, extra_demand), having the primary key (cell_id, time_stamp). The other database, CALL_BLOCK, stores the traffic history during the last update period, which is predefined. The database schema is CALL_BLOCK (cell_id, date, time_stamp, no_of_call_block), having the primary key (cell_id, date, time_stamp). It stores call block (if any) at different time stamps. The database EXT_DEM is updated with the help of the CALL_BLOCK database after each update period. To facilitate update, the database EXT_DEM is indexed on the primary key. In addition to the two databases, a central reserve pool (C_R_P) maintains a set of information

containing cell-wise assignment of extra demand. This information is related to extra demand assignment, namely EXT_DEM_ASSGN containing cell_id, extra_channel_assigned.

Periodic Update of EXT_DEM

In the database CALL_BLOCK, cell_id and time_stamp are maintained as constant while date varies. The attribute no_of_call_block measures the extra demand of channel at that time for that particular cell_id. The attribute extra_dem of the database EXT_DEM is automatically updated after every update period depending upon the average no_of_call_block for the last update period.

The Algorithm

The algorithm consists of two parts—the channel assignment algorithm and the channel release algorithm.

Channel Assignment Algorithm

The EXT_DEM database of each MSC triggers at each time slot, and accordingly requests are placed to the MSC for each cell's extra demand (if any). The MSC in turn runs the channel assignment algorithm. The following steps 1 to 4 are performed for the demands of each cell.

- **Step 1:** The EXT_DEM_ASSGN is consulted to know whether any of the co-channel cells of the demanding cell is assigned any channel, which is currently not being used by the cell. If so, the channel is assigned and a new entry is inserted in EXT_DEM_ASSGN.
- **Step 2:** Step 1 is continued until demand is exhausted or the channel used by co-

Figure 2. Percentage of call block vs. time

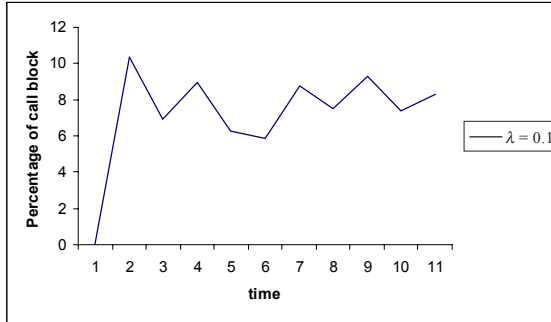
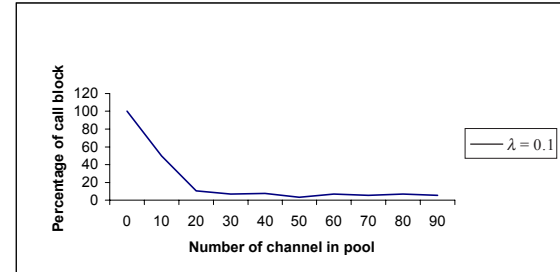


Figure 3. Percentage of call block vs. number of channel in pool



channel cells (which is not used by the demanding cell) is exhausted.

- **Step 3:** If demand is exhausted, control is sent to Step 4; otherwise new channels are assigned from C_R_P to the demanding cell, and a new entry in EXT_DEM_ASSGN is input for each of the assigned channels.
- **Step 4:** End of channel assignment for the current request.

Channel Release Algorithm

This algorithm runs for each request of releasing a channel.

- **Step 1:** The cell_id of the cell intending to release channel is found out.
- **Step 2:** The co-channel cells of that cell are identified.
- **Step 3:** The EXT_DEM_ASSGN is searched to find out whether any of the co-channel cells of the cell intending to release channel is using the channel to be released—that is, to find out whether there is any such tuple whose cell_id matches with any co-channel cell and extra_channel_assigned matches with the channel to be released.

- **Step 4:** If there is no such tuple, the channel may be declared as free and in turn the channel will be made available at C_R_P; otherwise it cannot be declared as free.

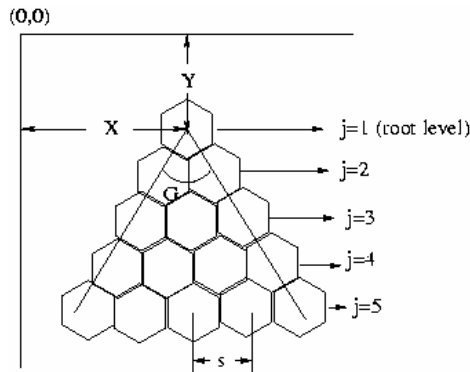
Simulation Results

In this section, performance of the scheme is presented. Detailed simulation experiments are carried out for the load balancing problem. Experiments are conducted to note the variation of the percentage of call block depending upon the time and number of channel in pool considering call arrival pattern that follows the Poisson process. Figure 2 shows the percentage of call block (PCB) as a function of time. It may be observed from the plot that the PCB reaches a steady state with time. Percentage of call block vs. number of channel in pool is shown in Figure 3. The result shows that after a sharp fall, the PCB reduces gradually.

Load Balancing Scheme II

The load balancing Scheme I does not consider the distribution of channels when a sudden known/unknown event occurs at any cell causing a sudden spurt of message traffic. Load

Figure 4. Cells of any group with $N=5$



balancing Scheme II (Mitra & Das(Bit), 2000) is a combination of dynamic channel assignment and channel borrow techniques. In this scheme, all the cells in the cellular mobile environment are divided into six groups, each group under one MSC. Channels are assigned dynamically among the cells under each MSC. In case of an unknown event for which channel demand may increase, the cell, which needs excess channels, can borrow a channel from one of the compact pattern cells within its own group (i.e., probable lender cell is described in a later section) by exchanging a negligible number of messages. If the probable lender cell is not available, the MSC can borrow channels from any other MSC that has excess channels. The whole scheme may be divided into three different sub-schemes: channel assignment scheme, channel borrow scheme, and channel release scheme. A few terms and ideas are defined here to explain the scheme.

Group Formation

The cellular architecture considers cells as hexagonal in shape. In such a configuration, for any cell there are 6 cells forming a ring around

it, and 12 cells surround this inner ring. These 12 cells may be considered as a next-level ring. So it may be generalized, as each ring contains $6*j$ cells where j is the level number, and $1 \leq j \leq N$ and N may be considered as the total ring number as well as the level number of a group. In the present work, cells of the system have been divided into six groups, with each group under one MSC. Each group, as shown in Figure 4, forms a tree-like structure with a cell in the root level (level 1) taken from ring 1. Level 2 of each group contains two cells from ring 2 and so on. So it may be generalized, as the structure has N number of levels. Thus, the number of levels depends on the number of rings considered.

Cell_id Calculation of the Cells in any Group

Considering that the rectangular coordinate of the root level cell of a group is (X, Y) , the cell_ids of the cells at any level of the same group can be determined as follows:

$$X\_coordinate = X + (N-1) * s * \cos(g * G + ((j-1) * (G/(N-1))))$$

$$Y\_coordinate = Y + (N-1) * s * \sin(g * G + ((j-1) * (G/(N-1))))$$

where 'g' is the group number ($1 \leq g \leq 6$). A group of five rings representing five levels is arranged as shown in Figure 4, where 's' stands for the distance between the centres of any two consecutive cells, and 'G' stands for the angle sustained at the centre of the root level cell by the distance between the centres of two successive corner cells.

Selection of Probable Lender Cell

The set of probable lender cells from which a hot cell can borrow channel can be selected as follows:

For a given cell let S_1 represent the set of CP cells and S_2 represent the set of cells that are in the same group with the given cell. Then $S_3 = S_1 \cap S_2$ represents the set of cells which are the CP cells of the given cell and also in the same group with the given cell. One of these cells can be selected as a lender.

Information Stored in MSC

The following information is stored at MSC to run the algorithm described in the next section.

- a. Cell_ids of all the cells under the MSC.
- b. Channel_ids of the allocated channels corresponding to each cell; number of channels (C) in each cell is computed by this information.
- c. Number of busy channels (B) in each cell; (C-B) indicating number of available channels (flag) in that cell; and (C-B)/C indicating degree of coldness is computed by this information.

If degree of coldness of a cell $\leq h$ (h is a threshold value > 0), the cell is designated as hot, otherwise it is treated as cold. Typical values of h are 0.2, 0.25, and so forth.

The Algorithm

The algorithm consisting of three parts—channel assignment, channel borrow, and channel release—is presented below.

Channel Assignment Algorithm

Whenever a call arrives, it is established through MSC. It keeps track of the number of calls established in each cell at any instance of time. To distribute the available channels among all the cells under an MSC dynamically, the MSC runs the following algorithm.

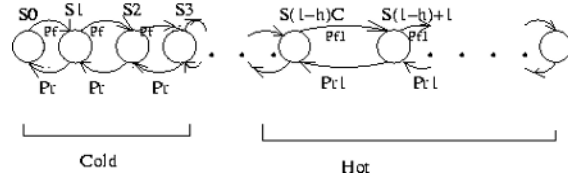
- **Step 1:** Initially a few channels are assigned to each cell.
- **Step 2:** The MSC assigns the total number of channels dynamically among the cells based on the call arrival probability in each cell.
- **Step 3:** When a channel is assigned to a cell, the number of busy channels (i.e., B) is increased by one; thus the number of available channels (i.e., the flag and degree of coldness) are updated accordingly.

Channel Borrow Algorithm

When a cell under an MSC becomes hot (i.e., $(C-B)/C \leq h$) due to the occurrence of any event, the MSC starts to run the channel borrow algorithm.

- **Step 1:** The X, Y coordinates (i.e., the cell_ids of all the CP cells of the hot cell) are computed.
- **Step 2:** From the CP cells computed in step 1 and the cells which are in the same group (as computed earlier) with the hot cell, probable lender cells are identified using the method discussed in the section “Selection of Probable Lender Cell.”
- **Step 3:** A list of probable lender cells from the identified cells computed in step 2 is prepared and ordered in descending order of their degree of coldness.
- **Step 4:** A check is performed taking the first element (cell having highest degree of coldness) from the list to find out whether the free channels available in the cell are being used by any of the co-channel cells. If not, and if the co-channel cells are not hot, free channel can be lent to the hot cell. Otherwise the same process is continued for the rest of the elements, one by one, until the hot cell gets a free channel. In case the demand of chan-

Figure 5. Markov model for a cell (Mitra & Das(Bit), 2000)



nel borrow is still not fulfilled, the MSC can borrow channels from other MSCs whose flag value is greater than a pre-defined threshold value.

- **Step 5:** As soon as a lender cell is selected, the co-channel cells of the lender cell are found out. The lent channel in all those co-channel cells is locked and the relevant stored data are updated accordingly.

Channel Release Algorithm

Whenever a cell intends to release a channel, MSC runs the channel release algorithm that contains the following steps:

- **Step 1:** The X,Y coordinates of the co-channel cells of the cell which intends to release channel are computed.
- **Step 2:** It is checked whether the channel to be released is currently being used by any of the co-channel cells computed in step 1. If it is not used currently, it can be declared as free, and the number of busy channels (i.e., B) is decreased by one; thus the number of available channels (i.e., the flag and degree of coldness) are updated accordingly.

Markov Modelling

In this section a discrete MARKOV model for a cell c in the system is derived. If the number

of occupied channel, say i, of the cell c is less than or equal to $(1-h) * C$ (h is threshold parameter and C is the total number of channel assigned to c), it may be considered as cold, else it is considered hot. In case of a cold cell (see Figure 5), probability (of being cold) P_i at any state S_i is given by $P_i = P_f * P_{i-1} + P_r * P_{i+1}$ (i) where P_f is the forward (cold to hot) transition probability and P_r is the reverse (hot to cold) transition probability. The solution of the equation (i) is as follows:

$$P_i = B * (P_f / P_r)^i \text{ (ii)}$$

Substituting the initial condition, i.e., $i=0, P_0 = B$. Hence $P_i = P_0 * (P_f / P_r)^i$ (iii)

Replacing (P_f / P_r) by l the equation (ii) becomes $P_i = P_0 * l^i$

The cell c becomes hot if the occupied channels of the cell c are in the range $(1-h)C$ to C.

For hot cell the state probability (of being hot) $P_i = B * (P_{fl} / P_{rl})^i$ (iv) where P_{fl} (hot to more hot) is the forward transition probability and P_{rl} (more hot to less hot) is the reverse transition probability. Substituting the initial condition (i.e., $i=(1-h)C$ in equation (iv)), the following is derived: $P_{(1-h)C} = B * (P_{fl} / P_{rl})^{(1-h)C}$ (v). On the other hand substituting i with $(1-h)C$, equation (iii) reduces to $P_{(1-h)C} = P_0 * (P_f / P_r)^{(1-h)C}$ (vi). From the equation (v) and (vi), we can deduce that $P_0 * (P_f / P_r)^{(1-h)C} = B * (P_{fl} / P_{rl})^{(1-h)C}$. Hence $B = P_0 * (P_f / P_r)^{(1-h)C} * (P_{rl} / P_{fl})^{(1-h)C}$. Replacing the value of B in equation (iv), the state probability $P_i = P_0 * l^{(1-h)C} * l_1^{-(1-h)C} * l_1^i$ where $(1-h)C \leq i \leq C$ and $l_1 = (P_{fl} / P_{rl})$. So the probability of a cell being hot is $P_h = P_i = P_0 * l^{(1-h)C} * l_1^{(i-(1-h)C)}$.

There is a probability of call block for the cell c if all of its channels are occupied. The cell is then in the state S_c . The call blocking probability is $P_c = l^{(1-h)C} * l_1^{hc} * P_0$.

Table 1. Messages exchanged to borrow a channel for one hot cell from a lender cell

Types of Messages	Communicated Between	Number of Messages	Delay
Request channel	MSC to lender BS	1	δ
	MSC to borrower BS	1	δ
Request locking	MSC to co-channel cells of lender cell	6	6δ

P_f defines the probability that the cold cell goes from state S_i to S_{i+1} . This transition can occur in any of the following ways:

- **Case(1):** There may be a new channel request in the cell with probability λ_1 .
- **Case(2):** A channel is lent on demand from another hot cell with probability λ_2 . This demand is satisfied only when there is no demand described in Case(1).
- **Case(3):** When there is neither Case (1) nor Case(2) type of demand, a cold cell can accept channel borrow demand from a remote hot cell with probability λ_3 . $P_f = \lambda_1 + (1-\lambda_1) \lambda_2 + (1-\lambda_1-\lambda_2) \lambda_3$.

Similarly P_r defines the probability that the hot cell goes from one state to another. This transition can occur in any of the following ways:

- **Case(1):** There may be a termination of call in the cell with probability μ .
- **Case(2):** Release of a lent channel. Probability of release of a lent channel is probability of local call termination (μ) + (probability of not termination of a local call $(1-\mu)$ * probability of termination of a call on lent channel (μ) * probability of lending a channel (λ_3). Hence $P_r = \mu + (1-\mu) \mu \lambda_3$.

For a hot cell the forward transition probability is P_{fi} . It occurs only when a new channel request comes. The probability is λ_1 . Hence $P_{fi} = \lambda_1$.

The reverse transition probability P_{ri} can occur for the following reasons:

- **Case(1):** There may be a local call termination. The probability is μ .
- **Case(2):** A channel borrow may occur. The probability of lending a channel is equal to the probability of borrowing demand from a hot cell—that is, λ_2 . Hence $P_{ri} = \mu + (1-\mu)\lambda_2$.

Now for given values of the parameters $\lambda_1, \lambda_2, \lambda_3, \mu, h$ and C , the values of P_f, P_r, P_{fi}, P_{ri} may be calculated. Substituting these values, l and l_i may be estimated, and finally the call blocking probability P_c may be determined empirically.

Message Exchange to Borrow a Single Channel

The performance of this scheme is analysed on the basis of the number of messages exchanged during channel borrow. The static links between all Base stations and the Mobile switching centre with a uniform message delay time δ is assumed. The messages and the correspond-

Figure 6. Percentage of call block vs. λ for various h

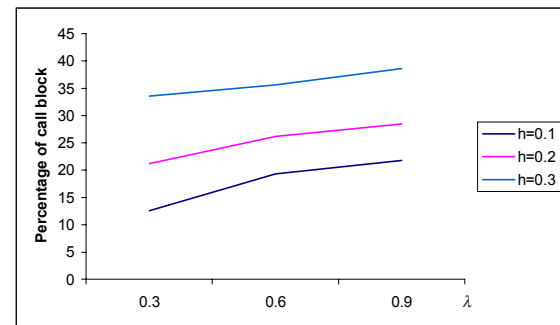
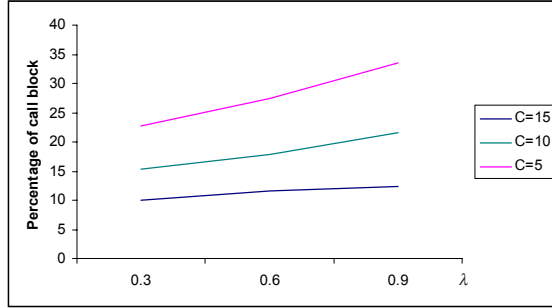


Figure 7. Percentage of call block vs. λ for various C



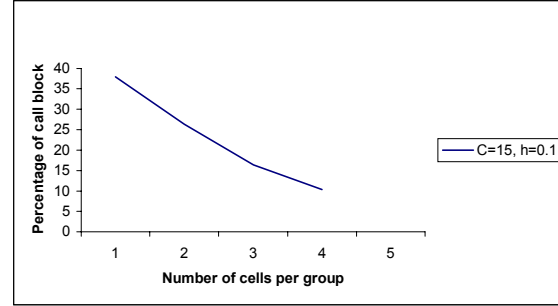
ing delay times of the proposed scheme are enumerated in Table 1. From Table 1 it has been observed that to borrow one channel for a hot cell, the total number of message exchanges is eight. If the probable lender cell is not available, the MSC which needs more channels can borrow channel from any other MSC with availability of excess channels by exchanging only two messages. In the worst case, the MSC, which needs excess channel, may send a request to the rest of five MSCs by exchanging 10 messages.

Simulation Results

In this section, the performance of the simulated system is presented. Detailed simulation studies are conducted to determine the percentage of call block as a function of call arrival probability λ considering call arrival pattern that follows the Poisson process. In one set of such experiments, PCB vs. λ is plotted for different values of the threshold parameter h . In the other set of experiments, the same is plotted, varying the number of channels assigned to any particular cell C .

Figure 6 shows the PCB as a function of call arrival probability for different values of threshold parameter h . As expected, it may be ob-

Figure 8. Percentage of call block vs. number of cells per group



served from this plot that the PCB increases with the increase in call arrival probability. Increase in threshold h leads to increase in the number of hot cells in the system. As a result, the number of cold cells from which channels can be borrowed reduces. Hence with the increase in h , the PCB increases.

Figure 7 shows the PCB as a function of call arrival probability for different values of C . It is to be noted that the PCB increases with the call arrival probability, but reduces with the increase in C as expected.

Finally, the PCB vs. number of cells in each group is plotted in Figure 8. It is observed that the PCB reduces with the increase in number of cells per group. With the increase in number of cells per group, the number of CP cells of the hot cell under the same MSC increases, which thereby increases the provision of borrowing channels, with a constraint that the maximum number of cells per group is much less than the total cells in the environment.

CONCLUSION AND FUTURE DIRECTION

In this chapter two load balancing schemes have been considered. Scheme I is appropriate

for steady state demand, whereas Scheme II is very much useful at the instant of occurrence of any known/unknown event. Scheme I may be enhanced keeping various event information in the database. The idea is EXT_DEM database now triggers not only at different time slots, but also whenever an event occurs, thus preventing any unnatural collapse of the system. Scheme II may be improved by forming the group dynamically so that the chances of getting probable lender cell are increased. In addition to this, both the schemes may be enhanced by incorporating a resource reservation mechanism in order to implement differentiated services. The users expecting better QoS have to pay more, whereas the users paying less get a best-effort service from the network.

As a concluding remark it can be said that now-a-days high-speed computers with huge storage are easily available and not much costly. On the other hand communication bandwidth is costly and not easily available. Hence the schemes discussed here emphasise reduction of message traffic, either with the help of database or employing critical computation resulting in the saving of bandwidth.

REFERENCES

- Boucher, N. J. (2001). *The cellular radio handbook*. New York: Wiley-Interscience.
- Das, S. K., Sen, S. K., & Jayaram, R. (1998). A novel load balancing scheme for the tele-traffic hot spot problem in cellular networks. *ACM/Baltzer Journal on Wireless Networks*, 4(4), 325-340.
- Das(Bit), S., & Mitra, S. (2000, September 25-27). Load balancing in a cellular mobile environment—A database approach. *Proceedings of the IEEE International Conference TENCON* (Vol. 2, pp. 195-200).
- Del Re, E. D., Fantacci, R., & Giambene, G. (1995). Handover and dynamic channel allocation techniques in mobile cellular networks. *IEEE Transactions on Vehicular Technology*, 44(2), 229-236.
- Lee, W. C. Y. (1995). *Mobile cellular telecommunications*. New York: McGraw-Hill.
- Macdonald, V.H. (1979). Advanced mobile phone service: The cellular concept. *Bell System Technical Journal*, 51, 15-44.
- Mitra, S., & Das(Bit), S. (2000). Load balancing strategy using dynamic channel assignment and channel borrowing in cellular mobile environment. *Proceedings of the International Conference ICPWC*, Hyderabad, India (pp. 278-282).
- Rappaport, T. S. (1999). *Wireless communications principles and practice*. Englewood Cliffs, NJ: Prentice-Hall.
- Tajima, J., & Imamura, K. (1988). A strategy for flexible channel assignment in mobile communication systems. *IEEE Transactions on Vehicular Technology*, 37(2), 92-103.

Chapter XXII

A Novel Fuzzy Scheduler for Mobile Ad Hoc Networks

S. Shanmugavel

Anna University, India

C. Gomathy

Deemed University, India

ABSTRACT

As mobile computing gains popularity, the need for ad hoc routing also continues to grow. In mobile ad hoc networks, the mobility of nodes and error prone nature of the wireless medium pose many challenges, including frequent route changes and packet losses. Such problems increase the packet delays and decrease the throughput. To meet with the dynamic queuing behaviour of Ad hoc networks, to provide QoS and hence to improve the performance, a scheduler can be used. This chapter presents a novel fuzzy based priority scheduler for mobile ad-hoc networks, to determine the priority of the packets. The performance of this scheduler is studied using GloMoSim and evaluated in terms of quantitative metrics such as packet delivery ratio, average end-to-end delay and throughput.

INTRODUCTION

A mobile ad hoc network is a cooperative engagement of mobile hosts or routers connected by wireless links. In the performance evaluation of a protocol, for an ad hoc network, the protocol should be tested under realistic conditions with representative data traffic models and realistic movement of mobile users. In order to thoroughly simulate a new protocol for

an ad hoc network, it is very essential to use a mobility model that accurately represents the mobile nodes (MNs). MNs within an ad hoc network move from location to location. A mobility model should attempt to mimic the movements of the real MNs. Currently, there are two types of mobility models used in simulations of ad hoc networks: traces and synthetic models (Camp, Boleng, & Davies, 2002; Lin, Noubir, & Rajaraman, 2004). Traces are those

mobility patterns that are observed in real-life systems. Traces provide accurate information when they involve a large number of participants and a long observation period, but privacy issues will prohibit the collection and distribution of such statistics, and new environments cannot be easily modeled. Hence, in these situations, synthetic models are used. They realistically represent MNs without the use of traces. We consider here three of the synthetic models—namely, random walk, random waypoint, and random direction mobility models (Bettsetter, 2001).

The random walk mobility model is a widely used mobility model and, in this, the current speed and direction of MN is independent of its past speed and direction. It has a memory-less mobility pattern, because it retains no knowledge containing its past location and speed values. Here, we encounter unrealistic generation of movements such as sudden stopping, sharp turning, and completely random wandering.

The random waypoint mobility model includes pause times between changes in direction and speed. An MN begins by staying in one location for a certain period of time (Jardosh, 2003; Camp et al., 2002). Once this time expires, the MN chooses a random destination in the simulation area and a speed that is uniformly distributed between minspeed and maxspeed. The MN then travels towards the newly chosen destination at the selected speed. Upon arrival, the MN pauses for a specified time period before starting the process again. This is also a widely used model. The RWP model is similar to the random walk model if pause time is zero.

The random direction mobility model is a revised version of random walk, and it ensures that every node is assigned the same speed throughout the entire simulation. After a random direction is chosen in the range 0 to 2π , an MN begins moving. If the MN reaches a grid

boundary, it bounces off the simulation border with an angle determined by the incoming direction. The MN then continues along this new path.

The choice of a mobility model can have a significant effect on the performance of an ad hoc network protocol. The performance of random walk, random waypoint, and random direction mobility models are compared. Dynamic source routing (DSR) protocol is chosen to be the routing protocol (Royer & Toh, 1999; Das, Castaneda, Yan, & Sengupta, 1998; Das, Perkins, & Royer, 2001). It determines the routes on demand. Here, the packet carries the full route that the packet should be able to traverse in its header.

DSR is chosen since it performs well in many performance evaluations of unicast protocols.

The performance metrics—namely, packet delivery ratio, end-to-end delay, average hop count, and protocol overhead—are used for comparison of these mobility models. The results prove that the random waypoint mobility model has the highest packet delivery ratio, lowest end-to-end delay, and lowest hop count (Camp et al., 2002). The random direction mobility model has the highest average hop count, highest end-to-end delay, and lowest packet delivery ratio since each MN moves to the border of the simulation area before changing its direction. The performance of the random walk model falls between these two. Hence to conclude, the random waypoint mobility model is used in many prominent simulation studies of ad hoc network protocols since it is flexible and it creates realistic mobility patterns for the way people might move in.

Research in the area of ad hoc networks has focused mainly on the routing protocols that decide the routing of packets hop by hop as efficiently as possible and medium access control (MAC), which indicates how to share the medium efficiently. But there is little focus

towards the queuing dynamics in the nodes of the networks and on the effects scheduling algorithms in the queues of the nodes. Hence, we believe that choice of scheduling algorithm will certainly improve the performance of the ad hoc network. Here, the different scheduling algorithms and the network's effect on mobile communication with the random waypoint mobility model are discussed.

We also design and analyze the performance of a fuzzy logic-based priority scheduler (FLPS), which combines the metrics and computes the crisp value of priority. The fuzzy algorithm for finding the priority of the packet based on some attributes of the packets is devised and coded in C language. The C code is linked with GloMoSim and is tested. It is found that the proposed fuzzy scheduler provides improved packet delivery ratio, reduced average end-to-end delay, and increased throughput when tested with various unicast routing protocols under different mobility conditions.

SCHEDULING ALGORITHMS

The ad hoc networks produce unique queuing dynamics due to the possible frequent transmission of control packets due to mobility, multi-hop forwarding of packets, and multiple roles of nodes as routers, sources, and sinks of data. The selection of the scheduling algorithms for mobile ad hoc networks is highly dependent on the queuing dynamics. These algorithms determine which queued packet to process next, and they have significant impact on the end-to-end performance. A scheduler for an ad hoc network is required to schedule the packets to reach the destination quickly, which are at the verge of expiry. The scheduler is positioned between the routing agent and the MAC layer. Without scheduling, packets will be processed in FIFO manner, and hence there are more chances that either more packets may be

dropped or may not meet the quality of service (QoS) target.

Generally, in all algorithms, high priority is given to control packets. Different drop policies are used for data and control packets when the buffer is full. When the incoming packet is a data packet, the data packet is dropped. If it is a control packet, the last enqueued data packet is dropped. If all packets are control packets, the incoming control packet is dropped.

There are many scheduling algorithms proposed in literature. No priority scheduling services both control and data packet in FIFO order. Priority scheduling gives high priority to control packets, and data packets are serviced in FIFO order. When considering the effect of setting priorities to data packets, these schedulers give high priority to control packets (Chun & Baker, 2002). Their differences are in assigning priorities among data queues. Weighted hop and weighted distance scheduling methods use the distance metrics. Weighted hop scheduling gives higher weight to data packets that have fewer remaining hops to traverse. If the packet has fewer remaining hops, then it has to reach the destination quickly. The data packets can be stored in round-robin fashion. The remaining hops to traverse can be obtained from packet headers. In weighted distance scheduling, physical distance is used. It is also a weighted round-robin scheduler. It gives higher weight to data packets, which have shorter geographic distances. The remaining distance is the distance between a chosen next hop and a destination. Round-robin scheduling maintains per-flow queues. The flow can be identified by a source and destination pair. Here each flow queue is allowed to send one packet at a time in a round-robin fashion. In greedy scheduling scheme, each node sends its own data packets before forwarding those of other nodes (Luo, Lu, & Bhargavan, 2000). The other nodes' data packets are serviced in FIFO order. Two other schedulers are earliest dead-

line first (EDF) and virtual clock (VC) (Kanodia, Li, Sabharwal, Sadeghi, & Knightly, 2002). In EDF, a packet arriving at time t and having delay bound d has a deadline $t + d$. In virtual clock, a packet with size L of a flow with service rate r has priority index L/r plus the maximum of current time t and priority index of the flow's previous packet.

In these priority scheduling algorithms we considered, the parameters used to find the priority of packets are: remaining hops to traverse, remaining distance, per-flow queues, greediness of nodes, delay bound, and service rate. With the thorough study of ad hoc networks, and the above mentioned scheduling algorithms, it is found that a number of metrics can be combined into a single decision so as to find the crisp value of the priority of packets. Our solution to determine the priority index of the packets utilizes the fuzzy logic concept (Gomathy & Shanmugavel, 2004). It deals with the imprecise and uncertain information of the network parameters since ad hoc network is dynamic in nature. This is advantageous in the target system because the fuzzy logic system is flexible and capable of operating with imprecise data and hence can be used to model nonlinear functions with arbitrary complexity. The fuzzy inference process works in three stages: fuzzification, rule evaluation, and defuzzification. In the first stage the parameters of the system are fed into a fuzzifier, which transforms the real-time measurements into fuzzy sets. The second stage applies a set of fuzzy rules onto fuzzy input in order to compute fuzzy outputs. Finally, outputs are translated into crisp values.

QoS Provisioning

QoS provisioning is becoming a critical issue in designing wireless ad hoc networks due to the necessity of providing multimedia applications in such networks. These are typically delay

sensitive and have high bandwidth requirements. It is a challenging task since the wireless channel is shared among adjacent hosts and network topology changes. Typical metrics for providing QoS include delay, loss rate bandwidth, and so forth. Here, in the design of scheduler, end-to-end delay and delivery ratio of packets are considered to analyze the performance of ad hoc networks and thus to provide QoS to the networks.

FUZZY LOGIC-BASED PRIORITY SCHEDULER (FLPS)

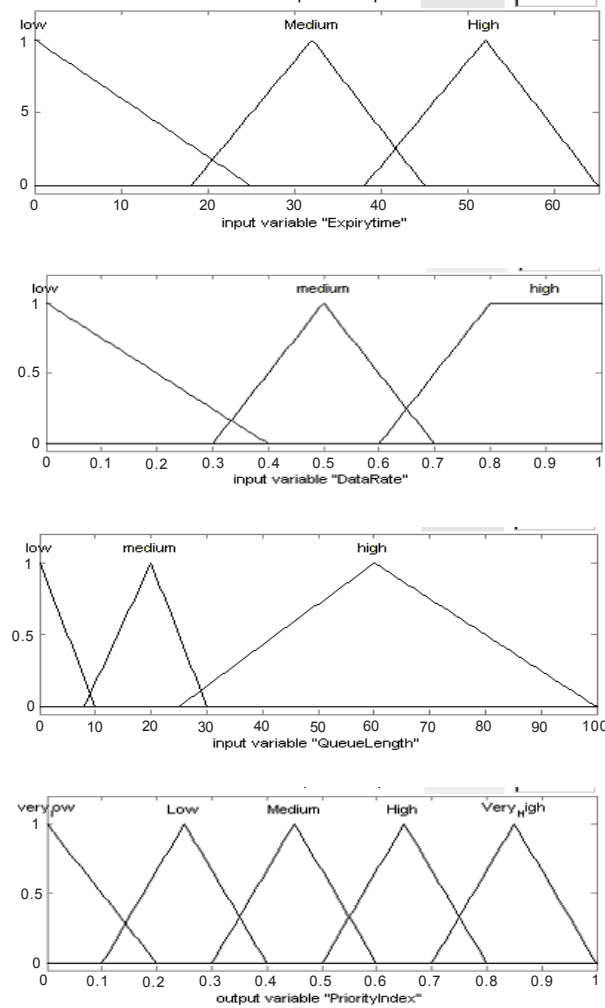
Fuzzy logic (FL) is a kind of artificial intelligence technology, which has the capability of mimicking how the field experts would make decisions. In networking area, there are a variety of traffic mixes, different traffic types, and long- and short-lived traffic flows. FL can incorporate networking expert knowledge to generate sensible solutions.

In order to improve the performance of scheduling algorithms, a mechanism using FL has been proposed to find the value of priority index of packets. The fuzzy logic-based priority scheduler aims to determine the priority index of data packets based on three input variables: data rate, expiry time of the packets, and queue length of the nodes. With these inputs, the fuzzification, rule evaluation, and defuzzification are performed. The following section describes the three processes.

Fuzzification

The three inputs are translated into fuzzy sets. The fuzzy sets contain elements that have a varying degree of membership in a set. Therefore, it is different from an ordinary set, where elements will only be considered as members of a class if they have full membership in that class. For example, if the expiry time is consid-

Figure 1. Membership functions of input and output variables



ered in an ordinary set, then it can only be either low or high and not both simultaneously, whereas in a fuzzy set, it can be classed as quite low, not so high, or medium and high. This indicates that the element in the fuzzy set can have membership in more than one set. The membership values are obtained by mapping the values obtained for a particular parameter onto a membership function, which will be used to determine the system outputs. This function is a curve or line that defines how each data or

value is mapped onto a membership value. We define here what are low L, medium M, and high H for each fuzzy set. Then by mapping the position of current input value onto the graph of the membership function, the input is allocated with a membership value in each set ranging from 0 to 1. Fuzzification of output parameter priority index is also performed, and five linguistic terms are attached to it. Figure 1 shows the membership function of the input and output variables for the FLPS.

Table 1. Fuzzy rule base (D =data rate; Q =queue length)

Q	L	M	H
D			
Expiry Time-Low			
L	L	L	VL
M	VL	VL	VL
H	L	VL	VL
Expiry Time-Medium			
L	M	M	L
M	M	M	L
H	M	M	M
Expiry Time-High			
L	VH	VH	H
M	H	M	M
H	H	H	M

Rule Evaluation

This stage involves feeding the fuzzy sets into an inference engine, where a set of fuzzy rules is applied. Fuzzy rules are usually defined as a set of possible scenarios in the form of if/then rules which determines the value of the priority index. Table 1 provides the summary of the decision-making logic. The first rule can be interpreted as: if the expiry time is low and data rate is low and queue length is low, then priority index is low. This indicates that, if expiry time of packets is low, it shows that packets are at the verge of expiry. And even if the data rate and number of packets in queue is low, the priority index is set to be low, so as to enable the packets to reach the destination quickly. The index value if very low indicates that packets are attached with high priority and will be scheduled immediately. If the index is very high, then packets are with lowest priority, and will be scheduled only after all high-priority packets are scheduled (Gomathy & Shanmugavel, 2004).

The rationale behind setting rules is that if the expiry time is low and both data rate and queue length are high, packets are attached with a very low priority index and hence possess high priority. Whereas if expiry time is high and both data rate and queue length are low, packets are attached with higher priority index

and hence possess low priority. These rules are then applied to fuzzy inputs and return the fuzzy outputs.

Defuzzification

At this stage the resultant fuzzy decision sets have to be converted into precise quantities. There exist several heuristic defuzzification methods such as max criterion, mean of maximum, and center of area or centroid method. In the FLPS, we consider the weighted average method of defuzzification to find the crisp output. The weighted average defuzzification technique can be expressed as:

$$x^* = \frac{\sum_i w_i \cdot m_i}{\sum_i w_i}$$

where x^* is the defuzzified output, m^i is the membership of the output of each rule, and w_i is the weight associated with each rule.

Example

Consider the scenario when the packets have an expiry time of 20 seconds, queue length of node the packet reaches is 50, and normalized data rate is 0.66 (normalized with respect to the channel capacity of 2 Mbps). For these set of inputs, the priority index is calculated as follows, which is done in three stages.

- **Fuzzification:** As seen from the figure, the expiry time of 20 seconds is fuzzified into low expiry with a degree of 0.2 and medium expiry with degree of 0.07. Queue length is fuzzified into high queue with degree 0.7142. Similarly, the data rate is fuzzified into medium rate with degree 0.2 and high rate with degree 0.3.

- **Rule Evaluation:** Now a series of if/then rules which are provided in the table are applied in order to determine the fuzzy outputs. An example of firing of rules is shown below.
 - If expiry time is low and data rate is medium and queue length is high, then priority index is very low.
 - If expiry time is medium and data rate is medium and queue length is high, then priority index is low.
 - If expiry time is low and data rate is high and queue length is high, then priority index is very low.
 - If expiry time is medium and data rate is high and queue length is high, then priority index is medium.

Since the rules are connected by an AND operation, we calculate the minimum function—that is, $\min \{0.2, 0.3, 0.7142\} = 0.2$, and we cut the fuzzy set very low of the output parameter priority index at this minimum level. Similar steps are done to determine the index of other rules. The four results of output overlap and they yield overall result.
- **Defuzzification:** The results are still a fuzzy set. Therefore we have to choose the representative crisp value for getting the final output. For this purpose, the weighted average method of defuzzification is used which yields a crisp value of $P = 0.175$. This value of P indicates that the packets are attached with high priority and will be scheduled immediately.

PERFORMANCE EVALUATION

The simulation for evaluating the proposed fuzzy scheduler is implemented using GloMoSim

library. First, the input variables used in fuzzy logic C code are identified. Then the calculated priority index is used for scheduling the data packet. By this way of scheduling, the packets that are about to expire or the packets in highly congested queues are given first priority for sending. As a result of this, the number of packets delivered to the client node, the end-to-end delay of the packet transmission, and the throughput improve.

The inputs to the fuzzy system are identified by a complete search of the GloMoSim environment. The input expiry time is the variable TTL, which is present in the network layer of the simulator. TTL stands for “time to live”. If the packet suffers excessive delays and undergoes multi-hop, its TTL falls to zero. As a result of this, the packet is dropped. If this variable is used as an input to the scheduler for finding the priority index, a packet with a very low TTL value is given the highest priority. Hence due to this, the dropping of packets experiencing multi-hops gets reduced. The next input to the scheduler is the data rate of transmission and it is normalized. The third input to the scheduler is the queue length of the node in which the packet is present. If the packet is present in a highly crowded node, it suffers excessive delays and gets lost. So, such a packet is given a higher priority and hence it gets saved.

The priority index is calculated with the inputs obtained from the network layer. This is then added to the header associated with the packet. Hence whenever the packet reaches a node, its priority index is calculated and it is attached with it. Each node has three queues. Each queue in the node is sorted based on the priority index, and the packet with the lowest priority index (i.e., packet with the highest priority) is scheduled next, when the node gets the opportunity to send. By this method of scheduling, the overall performance increases.

Table 2. Packets delivered—for Unicast protocols, with FLPS and WS

PACKETS DELIVERED - FOR UNICAST PROTOCOLS, WITH FLPS AND WS

Routing Protocol		Packets Delivered	
		FLPS	WS
	AODV	33155	21818
	DSR	34503	21676
	WRP	32183	28373

THROUGHPUT - FOR UNICAST PROTOCOLS WITH FLPS AND WS

Routing Protocol		Throughput	
		FLPS	WS
	AODV	263347	233513
	DSR	269890	164680
	WRP	263841	225320

END TO END DELAY - FOR UNICAST PROTOCOLS WITH FLPS AND WS

Routing Protocol		Average End to end delay	
		FLPS	WS
	AODV	0.97	1.127
	DSR	0.09	1.49
	WRP	0.3	0.571

Simulation Environment and Methodology

The simulation package GloMoSim is used to analyze and evaluate the performance of the proposed fuzzy scheduler. The GloMoSim (GLObal MObile information system SIMulator) provides a scalable simulation environment for wireless network systems. It is designed using the parallel discrete event simulation capability provided by PARSEC (PARallel Simulation Environment for Complex Systems) (Bargodia et al., 1999). It is a C-based simulation language developed by the parallel computing laboratory at UCLA (n.d.) for sequential and parallel execution of discrete event simulation model.

A network of mobile nodes is modeled and placed randomly within a 1000x1000-meter area. There were no network partitions through-

out the simulation. Each simulation is executed for 600 seconds of simulation time. Transmission range is chosen to 250 meters. Multiple runs with different seed values were conducted for each scenario, and collected data was averaged over those runs. A free space propagation model was used in our experiments. A traffic generator was developed to simulate CBR sources. Data sessions with randomly selected sources and destinations were simulated. Each source transmits data packets at a minimum rate of four packets/second and a maximum rate of 10 packets/second. The data payload was chosen to be 512 bytes/second.

Performance Metrics

The *packet delivery ratio* is the ratio of the number of data packets actually delivered to

Figure 2. Packet delivery ratio vs. number of nodes

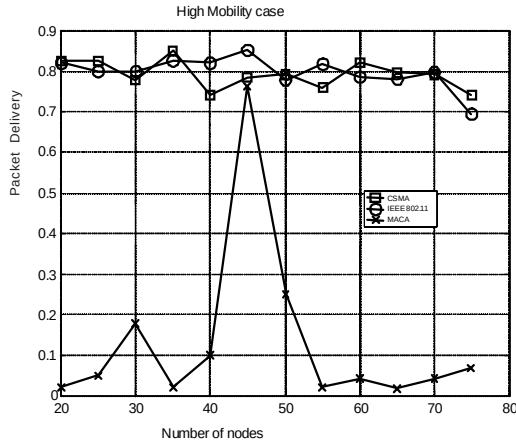
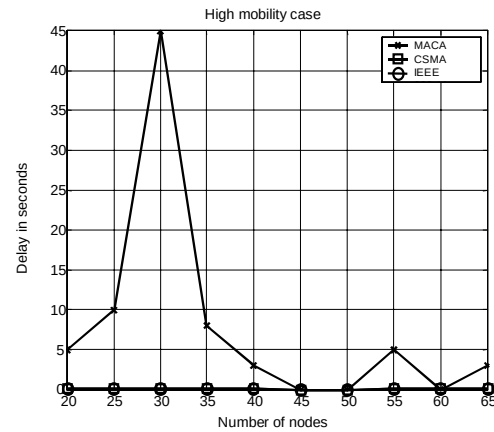


Figure 3. Average end-to-end delay as a function of number of nodes



the destinations to the number of data packets supposed to be received.

The *average end-to-end delay* indicates the end-to-end delay experienced by packets from source to destination. This includes the route discovery time, the queuing delay at node, the retransmission delay at the MAC layer, and the propagation and transfer time in the wireless channel.

Throughput is measured in bytes per second and serves as an effective performance metric.

The performance of the network with the fuzzy code (FLPS) and without the code (WS) is studied under and the routing protocols used in the simulator. The results are shown in Table 2 (Gomathy & Shanmugavel, 2004), showing that the proposed scheduler works well with the three routing protocols.

Scheduler Performance with Different MAC Layer Protocols: IEEE 802.11, CSMA, MACA

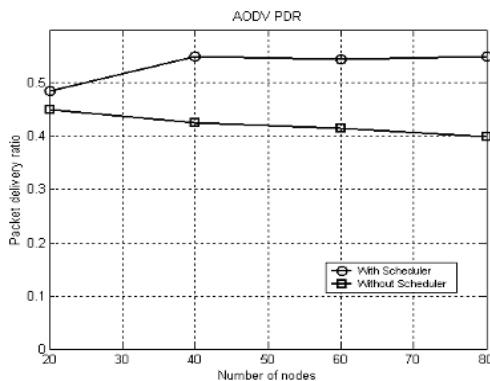
Experiments were performed to check the performance of the scheduler with different MAC

layer protocols such as IEEE 802.11, CSMA, and MACA. In the IEEE 802.11 protocol, each node maintains the scheduling table by overhearing all the RTSs and CTSs transmitted by other nodes within its broadcast range. Here an acknowledgment (ACK) of transmission is required after successful reception of data packet. In CSMA, if the transmission medium is in use, the node waits. It is limited by the hidden and exposed terminal problem. This can be solved by the use of RTS/CTS dialogue for collision avoidance. Hence IEEE 802.11 always shows a better performance as seen in Figures 2 and 3, compared to CSMA and MACA protocols (Gomathy & Shanmugavel, 2004). The collision avoidance mechanism in IEEE 802.11 aids in reducing the number of collisions, and hence more data packets reach the destination. Also in an exposed terminal scenario, both CSMA and MACA present poor performance behavior.

Scheduler Performance with Mobility: Two Nodes Transmitting at the Same Time to the Same Node

When two different nodes transmit at the same time to the same node with CSMA, less than half of the number of total packets is received by the receiving node due to collision. This scenario is presented for both MACA and IEEE 802.11 protocols. It is seen that a better behavior is obtained with IEEE 802.11. When used along with the scheduler, the performance with respect to throughput, packet delivery ratio, and delay improves further. The results are proved by experimenting with mobility changes under random waypoint condition in GloMoSim and are plotted. It is clear from the results that a fuzzy scheduler performs well with two nodes transmitting to the same node. Packet delivery ratio of the network with scheduler improves by 2-5%, as the mobility of the nodes varies from low to high range. The results are again verified for varieties of combinations of nodes, and results are averaged out (Gomathy & Shanmugavel, 2004). Similarly, there is an increase in throughput. There is a marked reduction in delay, which measures as low as 0.02 seconds under high mobility.

Figure 4. Packet delivery ratio as a function of network size



Variation in Network Size

In this simulation, the node mobility is set at 1m/s and network traffic load is made relatively heavy. The routing protocol is chosen to be AODV. Now the impact of node density on scheduler performance is studied. The packet delivery ratio is much improved as compared with that of one without scheduler, as seen in Figure 4. It is also seen that for lighter loads, the inclusion of the scheduler does not provide much improvement. But as traffic load increases, marked improvement is noticed in the performance. The average end-to-end delay performance proves that the end-to-end delay improves by 0.4 seconds when scheduler is included, as seen in Figure 5. This performance is achieved by the crisp calculation of the priority index including TTL as one of the inputs.

Variation in Mobility

In this simulation, each node is moved constantly with a predefined speed. The random waypoint mobility model is chosen for this study. The node movement speed or the mobility of the nodes is varied from 0 m/s to 10 m/s.

Figure 5. Average end-to-end delay vs. network size

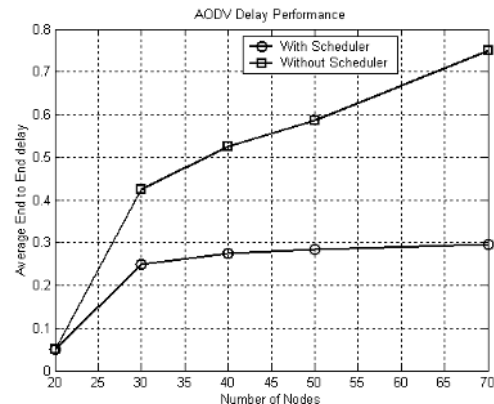
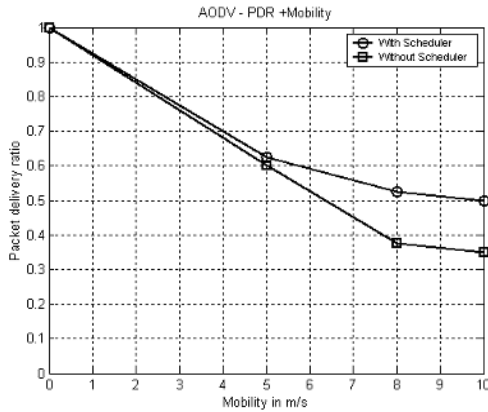


Figure 6. Packet delivery ratio vs. mobility

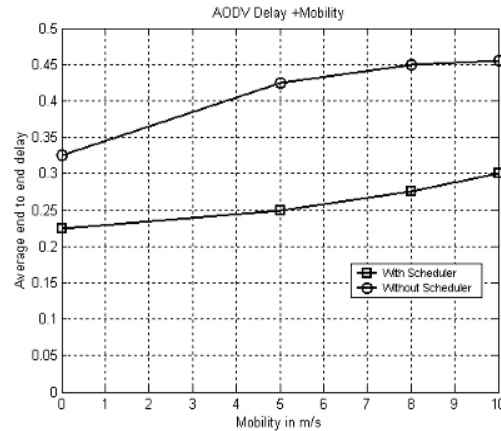


The number of nodes is set as 30 and the routing protocol is selected to be DSR. From the results in Figure 6, it is evident that the packet delivery ratio is at the higher side for the network with scheduler. Even though the delivery ratio reduces as mobility approaches 10 m/s, there is always an increase of 10% in the performance of fuzzy scheduler (Gomathy & Shanmugavel, 2004)

In the fuzzy scheduler, there is a slight degradation in performance as the number of nodes increases above 70. This is due to the increase in number of hops the packets have to take to reach the destination. But still, the end-to-end delay is much smaller compared to that of the network without the scheduler. It can be inferred from Figure 7 that the fuzzy scheduler provides a superior performance in terms of the end-to-end delay. As the mobility varies from 0-10 m/s, the fuzzy scheduler provides an end-to-end delay reduced by around 0.1 sec. to 0.2 sec. The performance of the fuzzy scheduler is also tested by varying the pause time, for RWP mobility model, using the AODV routing algorithm.

The results prove that the network performs better when FPLS is included. Increasing pause

Figure 7. Average end-to-end delay vs. mobility of nodes



time results in smoother transitions and hence improves the performance parameters. As the mobility model is changed to random walk, with uniformly distributed speed, the throughput packet delivery ratio reduces a little and the end-to-end delay increases slightly. This is because the random walk mobility model is characterized by the abrupt transitions in the direction and speed of nodes.

Scheduler Performance for Multicasting Protocols

Multicasting routing and packets forwarding in ad hoc networks is a fairly unexplored area. In today's network, data transmission between multiple senders and receivers is becoming increasingly important. Three multicast protocols are considered here for testing the proposed scheduler—ODMRP(On-Demand Multicast Routing Protocol), CAMP, and NTPMR (Node Transition Probability-based Multicast Routing) (Gomathy & Shanmugavel, 2005).

Variations in Mobility

In the mobility experiment, 20 nodes are multicast members and five sources transmit packets at the rate of two packets per second each. It is evident from the results that NTPMR provides higher packet delivery ratio as compared to ODMRP and CAMP. This is because NTPMR enables packets to travel distant destinations since a packet is sent to different neighbors during repeated encounters with a node. It is now proposed to include the fuzzy scheduler for these three protocols and test whether there is any improvement in packet delivery ratio. The packet delivery ratio (PDR) increases for all the three protocols. Hence it is verified that even at high mobility speeds, the routing protocols could be used when the fuzzy priority scheduler is included in these routing agents.

Multicast Group Size

The number of multicast members was varied to investigate the scalability of the protocol. The number of senders was fixed at five; the mobility speed at 1 m/s, network traffic rate at 10 packets per second, and the multicast group size was varied from 5 to 20 members. The routing effectiveness of the protocol as a function of multicast group size is now compared. For NTPMR, the packet delivery ratio is found to remain constant with increase in group size. Here the routing of packets does not depend on any forwarding group. CAMP performs better as the number of groups increases. Since the mesh becomes more massive with the growth of members, more redundant routes are formed. In ODMRP, as the number of receivers increases, the number of forwarding group nodes increases; this in turn increases the connectivity. With these results, the fuzzy scheduler is

inserted in between the MAC layer and routing agent. The simulation is run and the results are presented. As seen from results, the NTPMR shows an increased performance of 3%. This is again due to the fact that, with already existing best performance, as the data scheduler is added, the packets at the verge of expiry are scheduled immediately. This increases the PDR by 3%. For ODMRP, the scheduler PDR characteristics are closer to the one without scheduler. Again in CAMP, the PDR improves by 5% due to the proper selection of a priority index. Thus it is also verified that the proposed scheduler performs well with multicast protocols (Gomathy & Shanmugavel, 2005).

CONCLUSION

This chapter addresses a fuzzy-based priority scheduling scheme, which improves the quality of service parameters in mobile ad hoc networks. The fuzzy scheduler algorithm attaches a priority index to each data packet in the queue of the node. It combines the input parameters such as queue length, data rate, and expiry time to find the priority index. The crisp priority index is calculated by the fuzzy scheduler based on the above inputs, which are derived from the network. The membership functions and rule bases of the fuzzy scheduler are carefully designed. The coding is done in C language and output is verified using MATLAB fuzzy logic toolbox with FIS editor. Then the inputs are identified in the library of GloMoSim and the fuzzy scheduler is attached.

In this chapter, the performance of the fuzzy scheduler is studied for mobile ad hoc networks using GloMoSim simulator. It is found from the results that priority scheduling helps in effective routing of packets without much loss and with less delay. In a real network environment,

where timely reception of each packet plays a crucial role, priority scheduling helps in effective transmission of packets. Based on the studies, we conclude that the proposed fuzzy-based scheduling algorithm performs better compared with the network performance without scheduler. The results are also verified for different unicast and multicast routing protocols under different mobility conditions and network size, with IEEE 802.11 as MAC protocol.

The future extension of the work could be to include the mobility rate, number of nodes in the transmission range, channel state conditions, and fairness among sources as inputs to the fuzzy scheduler, and investigate the effect on the overall performance of the network.

REFERENCES

- Bargodia, R., Meyer, R., Takai, M., Chen, Y., Zeng, X., Martin, J., & Song, H. Y. (1999). PARSEC: A parallel simulation environment for complex systems. *IEEE Computers*, 31(10), 77-85.
- Bettsetter, C. (2001, July). Smooth is better than sharp: A random mobility model for simulation of wireless networks. *Proceedings of the 4th ACM International Workshop on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, Rome, Italy.
- Camp, T., Boleng, J., & Davies, V. (2002). A survey of mobility models for ad hoc network research, wireless communication and mobile computing (WCMC). *Trends and Applications*, 2(5), 483-502.
- Chun, B., & Baker, M. (2002). Evaluation of packet scheduling algorithms in mobile ad hoc networks. *ACM Mobile Computing and Communication Review*, 6(3).
- Das, S. R., Castaneda, R., Yan, J., & Sengupta, R. (1998). Comparative performance evaluation of routing protocols for mobile ad hoc networks. *Proceedings of the 7th International Conference on Computer Communications and Networks* (pp. 153-161).
- Das, S. R., Perkins, C. E., & Royer, E. M. (2001). *IEEE Personal Communications Magazine*, 8(1), 16-29.
- Gomathy, C., & Shanmugavel, S. (2004a, February 29-March 4). Fuzzy based priority scheduler for mobile ad hoc networks. *Proceedings of the 3rd International Conference on Networking*, Gosier, Guadeloupe.
- Gomathy, C., & Shanmugavel, S. (2004b, March). An efficient fuzzy based priority scheduler for mobile ad hoc networks and performance analysis for various mobility models. *Proceedings of the IEEE Wireless Communication and Networking Conference*, Atlanta, GA.
- Gomathy, C., & Shanmugavel, S. (2004c, December 11-14). Effect of packet scheduling and evaluation of fuzzy based priority scheduler on ad hoc network unicast communication. *Proceedings of the IEEE International Conference on Signal Processing and Communication*, Bangalore, India.
- Gomathy, C., & Shanmugavel, S. (2004d, December 15-18). Performance evaluation of a novel fuzzy based priority scheduler for mobile ad hoc networks and its effect on MAC protocols. *Proceedings of the 12th International Conference on Advanced Computing and Communication*, Ahmedabad, India.
- Gomathy, C., & Shanmugavel, S. (2005a, January 23-25). Design of a priority scheduler using fuzzy logic and the performance analysis with multicast routing protocols. *Proceedings of the IEEE International Conference on Per-*

sonal Wireless Communication, New Delhi, India.

Gomathy, C., & Shanmugavel, S. (2005b, January 28-30). Implementation of modified fuzzy priority scheduler for MANET and performance analysis with mixed traffic. *Proceedings of the 11th National Conference on Communication*, Kharagpur, India.

Gomathy, C., & Shanmugavel, S. (2005c). Performance evaluation of a novel fuzzy based priority scheduler for mobile ad hoc networks and its effect on MAC protocols. *International Journal of Information Technology*, 4(1), 78-86.

Jardesh, A., Royer, E. M., Kelvin, C., Almeroth, & Suri, S. (2003). Towards realistic mobility models for mobile ad hoc networks. *Proceedings of MOBICOM 2003*, San Diego, CA.

Kanodia, V., Li, C., Sabharwal, A., Sadeghi, B., & Knightly, E. (2002). Distributed priority scheduling and medium access in ad hoc networks. *ACM Wireless Networks*, 8(1).

Lin, G., Noubir, G., & Rajaraman, R. (2004). Mobility models for ad hoc network simulation. *Proceedings of IEEE INFOCOM 2004*.

Luo, H., Lu, S., & Bhargavan, V. (2000, August). A new model for packet scheduling in multi hop wireless networks. *Proceedings of ACM MobiCom'00*, Boston.

Rea, S., & Pesch, D. (2004, September). Multi-metric routing decisions for ad-hoc networks using fuzzy logic. *Proceedings of the 1st IEEE International Symposium on Wireless Communication Systems*, Mauritius.

Royer, E. M., & Toh, C. (1999). A review of current routing protocols for ad hoc networks. *IEEE Personal Communication*, 6(2), 46-55.

UCLA. (n.d.). *Parallel Computing Laboratory and Wireless Adaptive Mobility Laboratory: GloMoSim, a scalable simulation environment for wireless and wired network systems*. Retrieved from <http://pcl.cs.ucla.edu/projects/domains/glomosim.html>

Chapter XXIII

Co-Existence of WLAN and WPAN Communication Systems

Khaled Shuaib

United Arab Emirates University, UAE

Mohamed Boulmalf

United Arab Emirates University, UAE

ABSTRACT

Recently applications and technologies utilizing the free industrial, scientific, and medical (ISM) band have grown exponentially. Mainly there are three dominant technologies operating at the ISM 2.4 GHz band, IEEE 802.11 b/g, Bluetooth and IEEE 802.15.4 or Zigbee. With the diverse deployment and broad range of applications running over such technologies, it is inevitable that radio channel interference between devices utilizing such technologies exist. In this chapter we focus on co-existence issues between such technologies and on the quantification of the impact of Bluetooth on IEEE 802.11b/g.

INTRODUCTION

The advances achieved recently in the field of data wireless networks facilitated the introduction of wireless local area networks (WLANs) almost everywhere—in schools, hospitals, homes, office buildings, and even military bases. WLANs can be used to replace wired LANs or to extend their infrastructure, providing more

mobility and flexibility. Currently, not only laptops and desktops are equipped with WLAN interfaces, but many other devices such as PDAs, home appliances, and cell phones are as well. There are several WLAN standards developed by IEEE workgroups. The most widely used are the IEEE 802.11a, IEEE 802.11 b, and recently IEEE 802.11g. Table 1 summarizes technical aspects of these standards along with

other short-range wireless personal area network (WPAN) technologies such as Bluetooth and IEEE 802.15.4 or Zigbee.

There are many environments where WPAN devices can be found operating within the range of an installed WLAN network. An example of such environments could be a college campus, where students might be using Bluetooth to exchange information in an ad hoc fashion while others are connected to a WLAN network for Web browsing, taking an online exam, downloading files, or conducting office work. Another good example of a WPAN environment is a medical facility. In such an environment, many medical applications can be facilitated with the use of WLAN, Bluetooth, and Zigbee devices. For example, using their mobile devices equipped with a WLAN interface, physicians can electronically issue prescriptions while on rounds, immediately sending the prescription to the patient's choice of pharmacy while automatically checking for known allergies and conflicting medications. Another application is that physicians can electronically order detailed tests for their patients and securely view results as soon as they are available. As for Bluetooth and Zigbee, there are many applications that can apply in a medical facility. For example, Zigbee or Bluetooth wire-

less sensors can be used to monitor changes in patients' vital signs and alarm medical staff when needed. Moreover, Zigbee sensors in conjunction with WLAN can be used for location-aware services by automatically detecting the presence of a medical staff mobile device and updating it with needed records or updating the location server with the current medical staff location. Given the above examples, and there exist many more such examples, it becomes important to investigate the co-existence of WLAN and WPAN technologies within the same work space.

BACKGROUND

There has been a great interest among both industry and academia for research related to both WLAN and WPAN technologies. This section will briefly outline the characteristics of such technologies and discuss recent work done related to the co-existence of WLAN and WPAN networks.

IEEE 802.11

WLAN technologies as specified in IEEE 802.11 (1999) have been leading internal Internet dis-

Table 1. A comparison between the different WLAN and WPAN standards

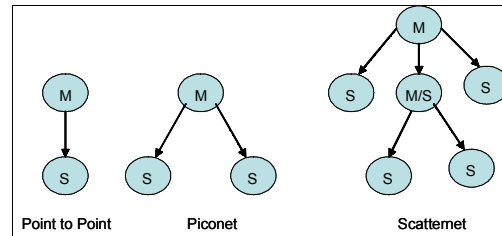
Technology	Frequency Allocation	Maximum Distance	Modulation Techniques	Maximum Throughput	Numbers of RF Channels
802.11a	5GHz	Up to 50 meters	COFDM BPSK; QPSK; 16 QAM	Up to 54Mbps	5
802.11b	2.4GHz	Up to 100 meters	QPSK (CCK)	Up to 11 Mbps	11
802.11g	2.4GHz	Up to 100 meters	OFDM 64 + Legacy CCK	Up to 54Mbps	11
Bluetooth	2.4GHz	Up to 30 meters	GFSK	720Kbps	79
Zigbee	2.4 GHZ 868-868.6 MHZ 902- 928 MHZ	Up to 30 meters	BPSK; OQPSK	250 Kbps; 40 Kbps	1; 10; 16

tribution in education, business, and home environments. These are usually deployed as the wireless extension of broadband access to a network. These technologies are based on CSMA/CA media access, with a positive MAC layer acknowledgment and a retransmission mechanism that aids noisy channel propagation condition and eventual undetected collisions. Today, WLAN standard defines high-rate data throughputs, such as the IEEE802.11b with a maximum throughput of 11Mbps and the IEEE802.11g with a maximum throughput of 54Mbps. Both 802.11b and 802.11g operate at the 2.4 GHz band. Typically, WLAN devices operate within 100 meters of distance range depending on the surrounding environment. While, the 802.11b utilizes direct sequence spread spectrum (DSSS) using complementary code keying (CCK) modulation, 802.11g is based on the orthogonal frequency division multiplexing (OFDM) modulation technique and the CCK modulation for backward compatibility with 802.11b.

Bluetooth

The Bluetooth (BT) standard (Bluetooth Special Interest Group, 2003) is a specification for WPAN. Although products based on the Bluetooth standard are often capable of operating at greater distances, the targeted operational area is the area around an individual (e.g., within 10 meters of the user). Bluetooth utilizes a short-range radio link that operates in the 2.4 GHz industrial scientific and medical (ISM) band similar to WLAN. However, the radio link in Bluetooth is based on frequency hopping spread spectrum techniques. Although at any point in time, the Bluetooth signal occupies only 1MHz, the signal changes center frequency (or hops) deterministically at a rate of 1,600 hops per second. Bluetooth hops over 79 center frequencies, so over time the Bluetooth signal

Figure 1. Bluetooth topology formations

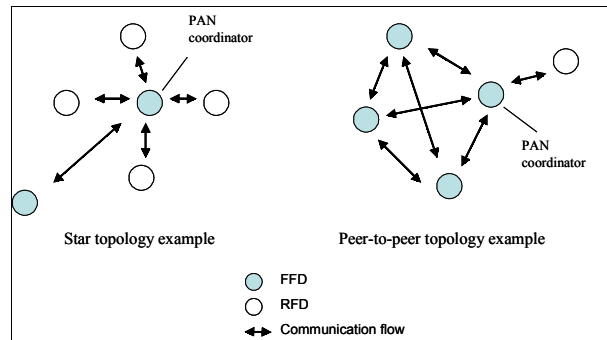


actually occupies 79MHz. Bluetooth is most suitable for small ad hoc network configurations, where devices can be operating either in a master or slave mode. With respect to topological formation of Bluetooth devices, several configurations are possible. For example, the configuration can consist only of two devices, one as a master and another as a slave as a point-to-point connection. Another setup is possible where up to seven devices can be clustered as slaves associated with one master device; this is called a Piconet. A more complex topology that permits the formation of larger networks would have several Piconets connected to form what is known as a Scatternet. Figure 1 shows the different possible topologies as specified by the Bluetooth standards.

IEEE 802.15.4

The new short-range, low-power, low-rate wireless networking Zigbee standard, IEEE P802.15.4/ D18 (2003), complements the high data rate technologies such as WLAN and opens the door for many new applications. This standard operates at two bands, the 2.4 GHz band with a maximum rate of 250 kbps and the 868-928 MHz band with data rates between 20 and 40 kbps. Zigbee is based on DSSS and uses binary phase shift keying (BPSK) in the 868/928 MHz bands and offset quadrature phase shift keying (O-QPSK) modulation at the 2.4

Figure 2. Zigbee star and peer-to-peer example topology



GHz band. While Bluetooth devices are more suited for fairly high-rate sensor applications and voice applications, Zigbee is better suited for low-rate sensors and devices used for control applications that do not require high data rate but must have long battery life, low user interventions, and mobile topology.

Zigbee is designed as a low-complexity, low-cost, low-power consumption, and low-data rate wireless connectivity standard. Zigbee supports scalable data rates; for example, it can be used for medical file transfer at the rate of 250 kbps while supporting sensor-based applications at the rate of 20 kbps.

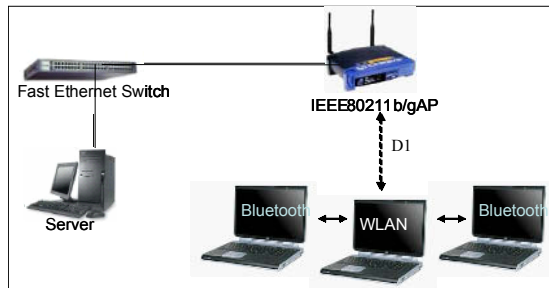
The Zigbee standards define two types of devices, a full-function device (FFD) and a reduced function device (RFD). The FFD can operate in three different modes—a personal area network (PAN) coordinator, a coordinator, or a device. The RFD is intended for very simple applications that do not require the transfer of large amount of data and need minimal resources. A WPAN is formed when at least two devices are communicating, one being an FFD assuming the role of a coordinator. Depending on the application requirements, Zigbee devices might operate either in a star topology or a peer-to-peer topology. Figure 2 shows both topologies. As seen in Figure 2, in a star topology the flow of communication is established

between devices and an FFD acting as the PAN coordinator. The PAN coordinator is a device that is responsible for initiating, terminating, and routing information around the network. The star topology is mostly used in small areas such as home automation, personal health care management, and hospital rooms, while the peer-to-peer topology is used in larger scale and more complex networks.

Co-Existence Issues

Devices operating at the 2.4 GHz frequency band have been growing due to the overwhelming applications these devices are being utilized for and the desire for ubiquitous networking. Examples of such devices are PDAs, cordless phones, home appliances, medical devices, mobiles, and laptop computers. With this increased use of devices at the 2.4 GHz band, the WLAN and WPAN devices are likely to be in close proximity to one another with possible interference. Therefore, recently several researches have started to investigate the mutual existence of such devices. For example, Golmie and Rebala (2003) investigated the use of adaptive techniques to mitigate interference for Bluetooth systems in the presence of WLAN direct sequence spread spectrum devices. Shoemake (2001) conducted a study on the

Figure 3. Basic test-bed for WLAN and Bluetooth throughput



interference between Bluetooth and IEEE 802.11b. Furthermore, the performance of different scheduling algorithms for Bluetooth intra-Piconet and inter-Piconets was evaluated by Har, Kofman, Segall, and Zussaman (2004), without examining the effect of a co-existing WLAN. Howitt and Gutierrez (2003) presented a brief technical introduction of the IEEE 802.15.4 standard and analyzed the coexistence impact of an IEEE 802.15.4 network on the IEEE 802.11b devices. The results presented were based on a theoretical analytical model and suggested that IEEE 802.15.4 will typically have little impact on the IEEE 802.11b performance.

In this chapter we present experimental results on the effect of Bluetooth existence on the performance of WLAN. At this time, there are no commercially produced Zigbee-enabled devices, and therefore no results are presented on the co-existence of Zigbee with WLAN or Bluetooth.

CO-EXISTENCE OF WLAN AND BLUETOOTH

As WPAN technologies become more integrated into devices used daily by almost every individual, and as WLAN becomes more domi-

nant all around the world, it is crucial to optimize such integrated deployment. This work involves the investigation of issues related to the co-existence of WLAN 802.11b/g and Bluetooth. Since Bluetooth devices hop over 79 MHz of the ISM band, it is not possible to have these products in the same area without the chance of interference. Due to the potential for interference, a series of coexistence tests needs to be run with actual products to determine their level of coexistence. Tests to be run are intended to obtain empirical throughput data corresponding with certain realistic scenarios in which WLAN and Bluetooth connections may coexist. It is important to realize that many different coexistence scenarios are probable in real-life usage, each with its own unique set-up characterized by different relative distances, applications, and performance measures. In this chapter we are more concerned with the effect of Bluetooth on WLAN, and therefore most of the results presented were to reflect that.

Test-Bed Description

Several tests were conducted to see the effect of Bluetooth on the performance of WLAN. These tests were performed in an open indoor (25m x 20m) cubical office environment area, with no interference from any other radio frequency devices except for the ones used in the test-bed. We used a Linksys 802.11 b/g access point (AP), two laptops with USP Bluetooth interfaces, and one laptop with an 802.11 interface. As seen from Figure 3, the AP and a PC server were connected to the Fast Ethernet switch. The laptop with WLAN was placed on a 0.5 meter high table a distance $D1$ from the base of the AP which was placed on a 2.5 meters wood post. UDP traffic was generated from a single source on the WLAN laptop and received at the server. The traffic was generated using the "LanTrafficV2" software with

packet payload size of 1,460 bytes and a fixed inter-packet delay of 1 ms. For all tests, 60,000 packets were transferred between the WLAN laptop and the server.

For all testing scenarios, the RTS/CTS protection mode was on at the 802.11b/g Linksys AP. In this mode, when a device wants to communicate, it sends a Request To Send (RTS) to the destination node, and waits for a Clear To Send (CTS) message before it transmits any data. This is done to avoid collisions, but it brings the maximum data throughput performance down (Broadcom, 2003). The RTS/CTS handshaking provides positive control over the use of the shared medium. The primary reason for implementing RTS/CTS is to minimize collisions among hidden stations. This occurs when users and access points are spread out throughout the facility and a relatively high number of retransmissions are occurring on the wireless LAN.

After a baseline testing of the throughput performance of WLAN 802.11g and 802.11b, the Bluetooth devices were introduced at different positions with respect to each other, the WLAN AP and the WLAN laptop. Figure 4 shows the general layout of the testing area, with the positions of each of the WLAN and

Bluetooth devices specified using the (x, y) coordinates for each test case. Figure 5 shows the baseline throughput performance of the 802.11g and 802.11b, for the following positions of the WLAN laptop with respect to the AP, (0, 3), (0, 16), and (0, 23). The signal strength for the taken positions as indicated by the Linksys AP software was at 100%, 94%, and 92%, respectively. The maximum obtained throughput for all test cases was reported as 5.5 Mbps for 802.11b and 11.3 Mbps for 802.11g, both at the WLAN laptop at the (0, 3) position with respect to the AP. The reported throughput for all test cases as measured by the "LanTrafficV2" was for the transport layer payload after the removal of all underlying headers.

Performance Results and Analysis

To look at the effect of Bluetooth devices on the performance of WLAN, two laptops were equipped with USP. Bluetooth interface cards were introduced into the test-bed area and placed at different positions with respect to both the WLAN AP and the laptop. All tests to investigate the effect of Bluetooth were run on the WLAN while a file was being transferred via FTP between the two laptops with Bluetooth.

Figure 4. Test-bed layout setup for WLAN and Bluetooth

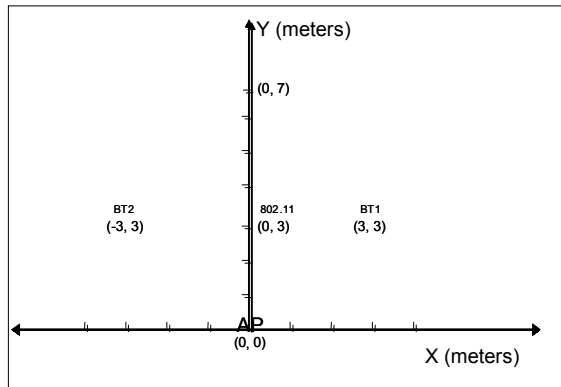


Figure 5. Baseline throughput results of 802.11g and 802.11b at positions (0, 3), (0, 16), and (0, 23)

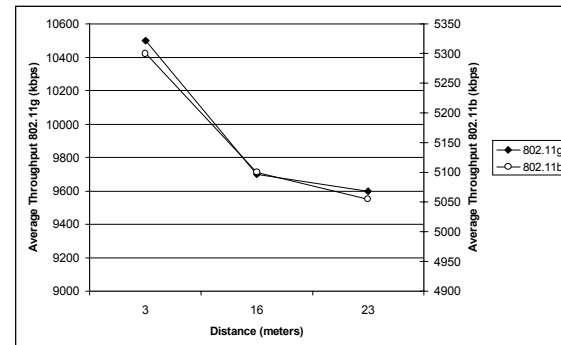


Figure 6. Performance of 802.11g in the presence of BT with the AP being 3m away from the 802.11 laptop

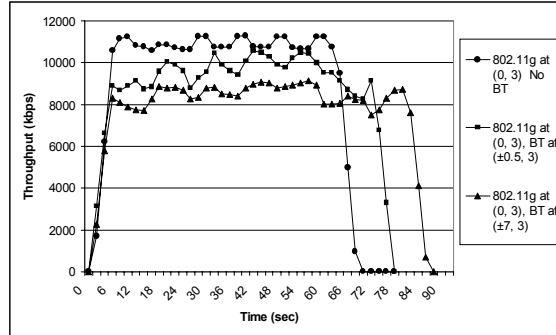
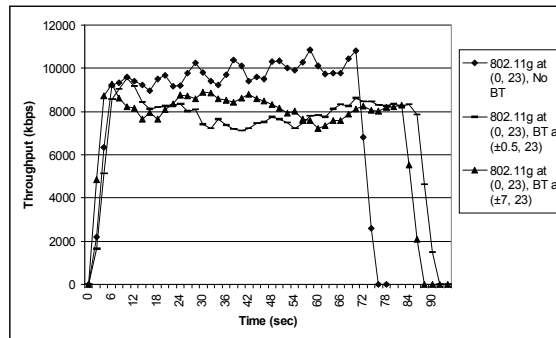


Figure 8. Performance of 802.11g in the presence of BT with the AP being 23m away from the 802.11 laptop



The Bluetooth devices used power control for all three specified Bluetooth device power classes (Bluetooth Special Interest Group, 1999). The Bluetooth Special Interest Group (1999) specification suggests that the transmitted power should be adjusted based on the received signal strength indicator measured at the receiver.

The first experiment was conducted to verify the effect of BT on WLAN when placed close to the AP with BT devices separated by a specific distance. In this scenario, the WLAN laptop was positioned at (0, 3), while the BT devices were placed at (0.5, 3) and (-0.5, 3)

Figure 7. Performance of 802.11b in the presence of BT with the AP being 3m away from the 802.11 laptop

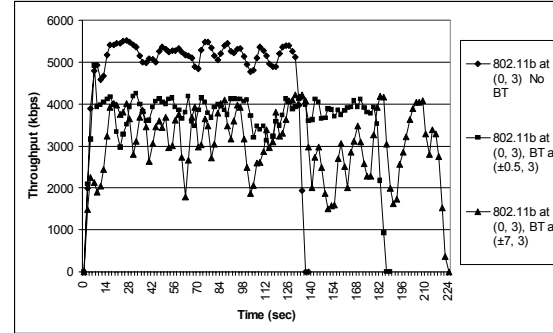
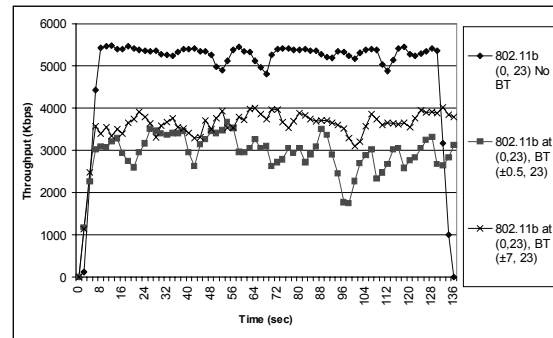


Figure 9. Performance of 802.11b in the presence of BT with the AP being 23m away from the 802.11 laptop



first, then at (7, 3) and (-7, 3). The WLAN throughput is shown in Figures 6 and 7 for the 802.11g and 802.11b respectively. As can be seen from the two figures, the BT devices greatly affected the WLAN throughput. An interesting observation is that as the distance between the two BT devices increased from 1m to 14m, the effect on the throughput of both 802.11b and 802.11g was worse. This is due to the power control mechanism used by the BT devices explained by the Bluetooth Special Interest Group (2003). For the 802.11b, the average throughput dropped by around 25% and 38% when the distance between the BT

Figure 10. Verifying the interference effect of the BT power control on the AP

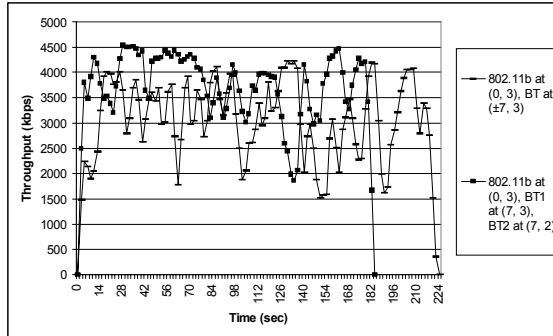
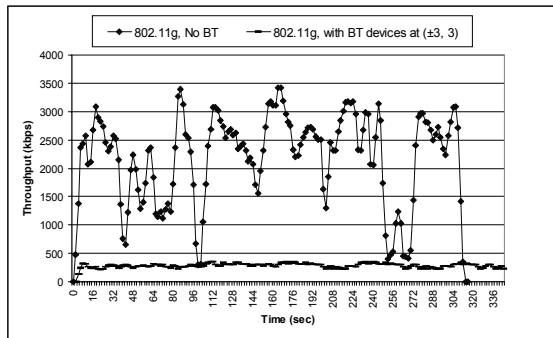


Figure 12. Performance results for 802.11g, with and without Bluetooth when the signal is at 40%



devices was 1m and 14m respectively. However, for the 802.11g the effect of BT was less drastic, with a drop of around 18% for the 1m separation and 22% for the 14m case.

To see the BT effect at a position considered far from the AP, the following positions were chosen: (0, 23) for the WLAN laptop, (± 0.5 , 23) and (± 7 , 23) for the BT devices. The results of this experiment are shown in Figures 8 and 9. As seen from these two figures, the achieved throughput was better for the case of BT devices positioned at (± 7 , 23)—that is, 14m apart with less effect on the 802.11g. The average throughput dropped by 43% and 30% for the 802.11b with BT at (± 0.5 , 23) and (± 7 ,

Figure 11. Performance results for 802.11b, with and without Bluetooth when the signal is at 40%

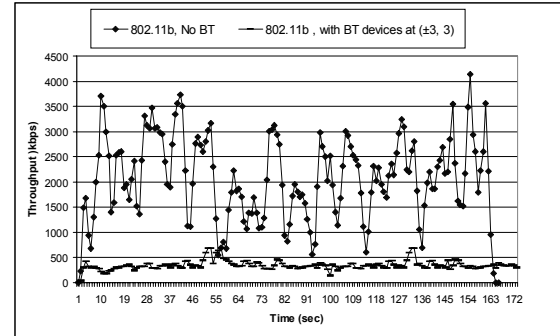
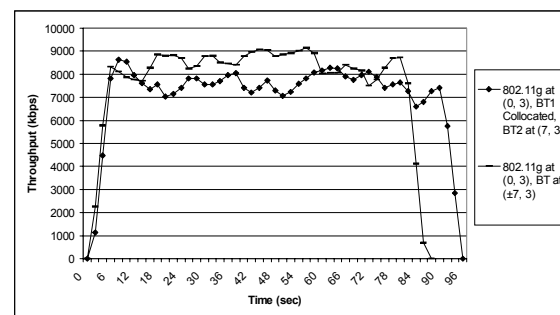


Figure 13. Results showing the Near Field Effect



23) respectively, while for 802.11g the results showed an 18% and 16% drop. These results are in line with the findings of the previous experiment, indicating that 802.11g is more immune to BT interferences than the 802.11b; however, they contradict the power control theory used to explain a better performance when the two BT devices are placed 1m apart rather than 14m apart. An explanation of this contradiction is that the BT devices' effect is mostly on the AP rather than the WLAN laptop interface card, and since the BT devices are for this latter experiment much farther away from the AP, the effect of the power control is minimal. To further analyze the power control

effect, another test scenario was performed where the BT devices were placed at (7, 3) and (7, 2), and the 802.11 laptop was kept at (0, 3)—that is, the BT devices were 1m apart, but 7m away from the 802.11 laptop. The results of this scenario were compared with those obtained when the BT devices were at (± 7 , 3). These results, shown in Figure 10, verify that the power control effect on the WLAN throughput is less as the distance between the BT devices was reduced from 14m to 1m while keeping the same coordinates with respect to the AP.

To see the effect of BT on WLAN, in an environment where the 802.11 signal strength is weak, we emulated positioning the WLAN laptop at a distance considered far from the AP. This was done by placing the AP behind an obstacle that brought down the signal strength indicator on the WLAN card software to 40%. At this signal strength level, a test was performed with all laptops being on the same horizontal line with the WLAN laptop at (0, 3) and the BT devices at (± 3 , 3). Figures 11 and 12 show the results of this experiment, comparing the WLAN throughput performance at this signal level with and without BT. In this test scenario the performance of 802.11g was revealed to be the same as that of 802.11b. Without any BT interference, the 802.11g average throughput was recorded at 2,200 kbps and that of 802.11b was 2,100 kbps. Once BT devices were introduced, the average throughput of 802.11g dropped by 87% to 282 kbps and that of 802.11b dropped by 83% to 349 kbps. This is due to the backward compatibility feature of 802.11g with 802.11b. Once the signal strength drops below a certain threshold, 802.11g, operates as an 802.11b using CCK modulation.

Certain laptops come with two PCMCIA slots collocated at the same place. For this test scenario, we placed an 802.11g card with a BT card together at the collocated PCMCIA slots

and looked at the effect of the introduced interference of BT on the performance of 802.11g. Figure 13 shows the throughput performance for this case when the 802.11g laptop was placed at (0, 3) and the second BT device was at (7, 3). The performance for this case was compared with that of having the 802.11g laptop at (0, 3) and the two BT devices at (± 7 , 3). By looking at the results, we notice that, when the BT device is collocated with 802.11g, the performance is worse than when the two BT devices are 14m apart. Considering the power control theory, this should not be the case; however, the results can be justified as valid due to the Near Field Effect theory. The theory states that a wireless device operating within the same frequency range of another should be collocated at a distance d so that: $d \gg 2D^2/\lambda$, where D is the length of the antenna on the wireless device and λ is the wavelength calculated as the speed of light divided by the operational frequency, 2.4 GHz.

CONCLUSION

The steep growth in the deployment of WLAN networks in office buildings, schools and colleges, medical facilities, army bases, shopping malls, hotels, and airports has been driven by increasing demand for wireless-based applications. This growth was also coupled with more advanced devices equipped with WPAN technologies such as Bluetooth. Yet another new standard for short-range wireless communications, IEEE 802.15.4 or Zigbee, was recently finalized and is currently being considered for ad hoc and sensor network applications. These technologies all operate at the same frequency band 2.4 GHz and are most likely to co-exist within the same work space. Therefore the potential for interference between devices operating these technologies within certain prox-

imity cannot be avoided. This chapter addressed performance issues related to the co-existence of WPAN devices within a WLAN network. Results presented showed how Bluetooth devices, when operating within the WLAN operational range, can greatly affect the throughput performance of WLAN devices.

REFERENCES

- Bluetooth Special Interest Group. (1999, December). *Specifications of the Bluetooth system* (vol. 1, v.1.0B 'Core' and vol.2.0B 'Profiles'). Retrieved from <http://www.Bluetooth.org>
- Bluetooth Special Interest Group. (2003, November). *Specifications of the Bluetooth system* (v.1.2). Retrieved from <http://www.Bluetooth.org>
- Broadcom. (2003). *IEEE 802.11g: The new mainstream wireless LAN standard*. White Paper. Irvine, CA. Retrieved from <http://www.broadcom.com>
- Golmie, N., & Rebala, O. (2003, December 1-5). Bluetooth adaptive techniques to mitigate interference. *Proceedings of GLOBE COM'03*, San Francisco (Vol. 1, pp. 405-409).
- Har, L., Kofman, R., Segall, A., & Zussaman, G. (2004). Load-adaptive inter-Piconet scheduling in small-scale Bluetooth Scatternets. *IEEE Communication Magazine*, (July), 136-142.
- Howitt, I., & Gutierrez, J. A. (2003, March 16-20). IEEE 802.15.4 low rate—wireless personal area network coexistence issues. *Proceedings of Wireless Communications and Networking 2003*, New Orleans, LA (Vol. 3).
- IEEE 802.11. (1999). *802.11 wireless medium access control and physical layer specifications*. Retrieved from <http://grouper.ieee.org/groups/802/11/>
- IEEE P802.15.4/ D18. (2003, February). *Low rate wireless personal area networks*. Retrieved from <http://grouper.ieee.org/groups/802/11/>
- Shoemake, M.B. (2001). *Wi-Fi (IEEE 802.11b) and Bluetooth coexistence issues and solutions for the 2.4 GHz ISM band*. White Paper. Dallas, TX: Texas Instruments. Retrieved from <http://www.ti.com>

Chapter XXIV

Mobility Support Resource Management for Mobile Networks

Mohammad Mahfuzul Islam
Monash University, Australia

Manzur Murshed
Monash University, Australia

ABSTRACT

Limitation of wireless resources is the main obstacle for meeting the widespread demands in the cellular network technology. The crisis of resources is further augmented when reservation is made for supporting the migrating users from the neighbouring cells. Many mobility support advanced technologies and mobile communications protocols have been developed to optimally utilize wireless resources. Some policies support the heterogeneous access technologies for multimedia services in mobile networks. Some other policies exploit the mobility information from the current and neighbouring cells to dynamically adjust the key components of resource management such as resource reservation, resource allocation, and call admission control policy to adapt quickly with network traffic changes. This chapter provides a comprehensive overview of various methods to explain the mechanisms for managing these key components of resource management for cellular networks.

INTRODUCTION

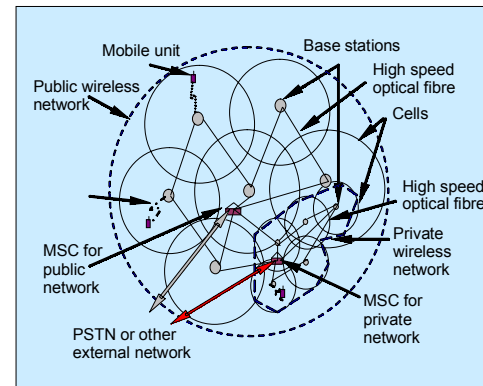
Next-generation mobile networks will support real-time interactive multimedia applications such as interactive video-on-demand, videoconferencing, teleconferencing, remote diagnosis, tele-operations, and so forth, besides its existing applications of voice conversation,

virtual office management, file downloading, e-mail, and Internet browsing. Mobile users are likely to demand the same *quality of service* (QoS) guarantees for multimedia data transmission as is available for wired networks. Currently, this expectation cannot be achieved, as the transmission capacity in wireless networks is much less than that in the wired

networks. Unlike wired networks, where it is possible to de-route the data or use more than one physical link to get higher capacity, wireless communication does not lend itself to such de-routing. Furthermore, mobility of hosts and channel imperfection make the QoS provision a far more challenging task in mobile networks. For example, a mobile unit admitted into the networks in a cell with satisfactory resources may face difficulty in continuing with required QoS soon after handing it off into a cell having little or no resources to offer. The problem becomes even more challenging with next-generation wireless networks where smaller-size cells will be implemented to allow higher transmission rates required for the multimedia applications. Smaller-size cells increase the handoff rate and result in rapid changes in the network traffic conditions, making the QoS guarantees more difficult (Oliviera, Kim, & Suda, 1998). So, resource management is the key area of research for getting effective implementation of mobile networks.

The bandwidth limitation problem in mobile networks can be partially compensated through transmitting data via wired medium, if possible. Therefore, a B3G/4G mobile network comprises both a wired network and a cellular wireless network. The *mobile unit* (MU), the base station, and the *mobile switching centre* (MSC) are the three constituent elements of a cellular network. An MU may be any handheld device, such as a mobile phone, portable computer, personal digital assistant (PDA), car communication systems, notebook, or any other device capable of communicating via omnidirectional radio waves using a prescribed communication protocol. The geographical area covered by a mobile communication network is divided into several regions known as cells, as shown in Figure 1 (Islam & Murshed, 2004). The communications between all the MUs inside a cell is controlled by a control centre

Figure 1. Typical heterogenous wireless network architecture



known as the base station through using one or more antennae connected with it. To facilitate the communications beyond the coverage of a single cell, all the base stations within a large geographical area are controlled by an MSC. The MSC stores the relevant information associated with all the MUs under its control and acts as a gateway to the external networks such as the Internet, the public-switched telephone network (PSTN), or any other MSC. The base stations are connected to an MSC by the high-speed wired network. As the high-speed optic fibre is assumed to be used in the wired part of mobile networks, the resource crisis is concentrated in the cellular part. The resources in this chapter refer to the bandwidth of omnidirectional radio waves used in cellular networks for performing all sorts of wireless communications required for mobile business.

In the next-generation B3G/4G heterogeneous cellular network configurations, the cell size may vary from a few meters to few kilometres. The smaller cells may reside entirely or partially within a larger cell to meet the higher capacity demand in “hotspots” (congested areas). Besides using the cellular networks in public domain, smaller cellular net-

works including its own MSC will also be established in private domain to fulfil the needs of a specific private organisation, as shown in Figure 1. Moreover, in order to reduce the cost and provide superior management, several Pico-cells (radius < 100 meters), each containing their own antennae, will be grouped together to form a cluster, controlled by a single control centre.

The conventional global system for mobile communications (GSM) based cellular network supports only voice communications. The recent 2.5G/3G mobile networks partially implement multimedia applications such as e-mail, Web-browsing, file-downloading, and information required for office management activities. However, the B3G/4G mobile networks presented in Figure 1 will support all sorts of multimedia applications required for mobile business, personal entertainment, and office management.

The mobile multimedia services can be divided into two large groups: the Class I (C1)—real-time multimedia applications; and the Class II (C2)—non-real-time data traffics (Oliviera et al., 1998). The C1 traffics include interactive image, voice, video, and audio services. Most of the C1 services can tolerate the bandwidth fluctuations while transmitting from the users equipped with an adjustable codec (e.g., Kanakia, Mishra, & Reibman, 1996; Vickers, Lee, & Suda, 1997). In the period of high traffic loads, the users of these services can adjust the coding rate such that the QoS received at the destinations is still acceptable—that is, above a predefined threshold. These services are known as *adaptive* or *elastic* services. However, some of the C1 services are *non-elastic* and cannot adapt with bandwidth fluctuations. The C2 services, on the other hand, include non-real-time data such as e-mail, text and multimedia messaging, file transmission, and other TCP/IP traffics. In case of network congestions, the

C2 services can be buffered at a network node (e.g., base station) or at a user's station and transmit at a slower rate. There is no minimum required bandwidth—that is, the minimum required bandwidth is zero for the C2 services, as they can tolerate relatively large delay and bandwidth variations. Therefore, these services are known as *fully adaptive* or *fully elastic* services.

In a cellular network, when an active MU (i.e., MU on service) moves from one cell to another having little or no resource available to support and continue the service, the communications must be dropped. From the user's perspective, dropping an ongoing call is far more undesirable than blocking a new call. In order to keep enough resources to support the migrating services from the neighbouring cells, the base station of a cell needs to reserve resources, even if denying MUs requesting admission into the network. Reserving resources eventually decreases the dropping possibility of migrating services (i.e., *call dropping probability*, or CDP), but increases the blocking rate of newly originated services known as *call blocking probability* (CBP). As the designers' aims are to minimize both the CBP and the CDP, it is necessary to estimate the amount of resources to be reserved in each cell as accurately as possible so that the CDP is kept below a certain level with the minimum increase in CBP. On the other hand, reserving resources to support handoff services ultimately reduces resource utilization, as the reserved resources become wastage in the cell where the given MU does not move into. Accurate estimation of resources to be reserved in a cell, therefore, increases the network resource utilization.

The conventional resource management schemes either reserved a fixed amount of resources along with maintaining buffers (i.e., queues) in each cell (e.g., Gaasvik, Cornefjord, & Svensson, 1991; Hong & Rappaport, 1986)

or used some assumptions about the handoff-probability through considering the MU's mobility as an random motion (e.g., El-Kadi, Olariu, & Abdel-Wahab, 2002; Kwon, Choi, Bisdikian, & Naghshineh, 2003; Hou & Fang, 2001; Fang & Zhang, 2002). The assumptions used in these schemes are as simple as reserving some resources for providing the higher priority to the handoff services such that the amount of resources that need to be reserved is estimated based on some blind experiences or statistical measurements. A more accurate amount of resource reservation requirements estimation is possible through considering the mobility information (e.g., historical databases, movement patterns, mobility parameters, etc.) directly as presented in some recent schemes (e.g., Islam & Murshed, 2004; Aljadhari & Znati, 2001; Levine, Akyildiz, & Naghshineh, 1997; Choi & Shin, 2002). These recent schemes estimate the probability of the MU to visit a given neighbouring cell using the mobility prediction techniques; form a cluster of neighbouring cells called *shadow cluster*, each of which is highly likely to be visited by the MU; and reserve the amount of resources in each cell of the shadow cluster based on the predicted mobility information.

Designing a *call admission control* (CAC) policy is another important area of resource management in cellular networks. Due to the limitation of resources, it is not possible to honour all the new and handoff services requesting admission into the cell. CAC policy decides whether an incoming call (new or handoff) is allowed to be admitted into a given cell for network service or not. When a new or handoff service gets admission into a cell of the network through CAC policy, the service may again suffer from getting its fair share of resources among the presence of the competing services in the cell. For an instant, some services may occupy the large amount of re-

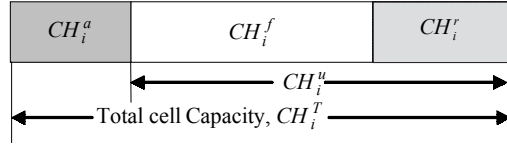
sources while the others are getting very little to survive. So, for effective and efficient wireless resource management in mobile networks, it is essential to design the appropriate CAC policy as well as the technique of fair resource distribution among the competing services, besides developing an accurate resource reservation method.

The aims of this chapter are to create a strong background for designing the effective and efficient protocol, the backbone, and the infrastructure of the next-generation B3G/4G mobile networks. To provide the complete views of resource management to both the developers and the researchers, this chapter briefly explained all of the key areas of mobile network resource management policies including: (1) mobility prediction and formation of the shadow cluster, (2) techniques of estimating the amount of resources to be reserved, (3) CAC policies, and (4) fair distribution of resources among the competing services in a cell. The methods presented in this chapter are suitable to be implemented in any of the standard protocols presented in a data-link layer, including the MAC protocol used in IEEE 802.11, Mobile-IP, and code division multiple access (CDMA) protocols.

FIXED OR STATISTICAL ASSUMPTIONS-BASED RESOURCE RESERVATION AND CALL ADMISSION CONTROL SCHEMES

For the last two decades, many research works have been proposed to optimally utilize the scarce resources of cellular networks for multimedia applications required in mobile business. For most of them, the CAC policy is designed through either reserving a fixed amount of bandwidth (Gaasvik et al., 1991; Hong & Rappaport, 1986) or estimating the CDP from

Figure 2. Distribution of cell capacity



the statistical assumptions of having an equal handoff probability to each of the neighbouring cells and keeping it less than a predefined threshold (see Wu, Wong, & Li, 2002; Kown, Choi, Bisdikian, & Naghshineh, 2003). These schemes, however, do not use any mobility information directly, but give the higher priority to the handoff services through reserving resources for them or restricting the new services to enter into the network in the congested period.

The conventional resource management scheme, known as the guard channel (GC) scheme, reserves a fixed portion of bandwidth—that is, channels in each cell exclusively for the handoff services (see Gaasvik et al., 1991; Hong & Rappaport, 1986; Fang, Chlamtac, & Lin, 1998; Posner & Guerin, 1985). The remaining unused channels are shared by both the handoff and the new services. Assume the cell i has the bandwidth capacity of C_i^T units which is divided into CH_i^T equal channels. Let CH_i^u be the number of unused channels in cell i found by subtracting the already allocated channels CH_i^a from the total cell capacity CH_i^T . CH_i^r is the number of unused channels reserved for handoff services, as shown in Figure 2. According to the GC scheme, a new service is initiated if the remaining unused channels, called free channels (i.e., $CH_i^f = CH_i^u - CH_i^r$) can support it, and a handoff service is continued as long as it can be supported by the total unused channels CH_i^u .

Due to the popularity of the GC scheme, numerous extensions have been proposed in literature. Ramjee, Nagarajan, and Towsley

(1996) proposed the fractional GC policy to minimize the CBP subject to a hard constraint on the CDP. Acampora and Naghshineh (1994) suggested forming a cluster comprising of a group of neighbouring cells in which a common reserve pool is maintained through allocating a portion of channels from each cell for the handoff services. Li, Lin, and Chanson (1998) extended the idea of the GC scheme to accommodate multiple traffic streams having potential differences in the QoS requirements and channel thresholds. Hou and Fang (2001), however, modified the GC scheme where, instead of reserving a fixed amount of resources, an estimated number of channels are reserved to provide some degree of dynamism. From the simulation study (see Hou & Fang, 2001), we find that the GC scheme with an estimated amount of channel reservation outperforms the GC scheme with fixed reservation or the fractional GC scheme. The GC scheme and its modifications are very simple because no communication and computational overheads are involved. All such schemes, however, are static and cannot adapt the changes in traffic conditions observed in real-time wireless networks, since these schemes do not use any traffic information from the current or the neighbouring cells.

To reduce the CDP further, Hong and Rappaport (1986) introduced the hand-off service queuing (HQ) model and used it along with the GC scheme to support the migrating (i.e., handoff) services when the channels are not available in the migrating cell. The fundamental assumption of the HQ model is that an MU can communicate simultaneously with more than one cell, and handoff decision is taken based on the received signal strength indicator (RSSI) level. If the RSSI level for an MU in the base station of the current cell is less than the handoff threshold signal strength level and also less than the RSSI in any neighbouring cell's base station, then the handoff procedure is

implemented. If there is no channel available in the next cell, the service is queued until a channel becomes available. However, if the RSSI drops below an acceptable received threshold, then the service is dropped. The performance of the HQ model has also been studied experimentally (see Gaasvik et al., 1991) and analytically through using the birth-death queuing model (see Van & Glisic, 2001). The conventional HQ model uses regular first-in-first-out (FIFO) queue without maintaining any priority scheme among the queued services. Tekinay and Jabbari (1992) extended the HQ model to a measured based priority-queue scheme (MBPS) that maintains the priority among the services queued based on the RSSI level—that is, higher priority is given to the services having lower RSSI level. When a channel becomes free in the migrating cell, the services in the queue having minimum RSSI level are granted access to that channel irrespective of their arrival time to the queue. McMillan (1995) proposed a dual buffer priority HQ model in which two different priority queues are maintained for the new and the handoff services respectively. All these schemes, however, are not suitable for multimedia communications for the reason that a huge volume of data cannot be stored in queues. Moreover, channels must be available before time expires, which is not guaranteed in the HQ model.

Some schemes (e.g., Wu et al., 2002; Kown et al., 2003; Fang & Zhang, 2002) consider the symmetric random walk model (SRWM) to predict the steady state distribution of users' demand for regulating the admission of new and/or handoff services with the aims of reducing the CDP. The SRWM assumes that the MU moves randomly in different directions at different speeds and hands off to any one of the neighbouring cells with equal handoff probability. Most of the SRWM-based schemes estimate the CDP using the different statistical

measurements so that the admission decision can be taken for the new services based on the criteria of keeping the CDP below a predefined threshold. Wu et al. (2002) used diffusion equation (see Kampen, 1981) to represent the movement of the MU and solved the equation to get the CDP for the CAC decision-making process. A two-dimensional Markov chain model is used in some schemes (e.g., Hou & Fang, 2001; Fang & Zhang, 2002) to estimate the CBP and the CDP in a given cell based on the assumption that the number of new services in a cell must be bounded by a given predefined number. The SRWM-based CAC schemes, however, consider the number of MUs in the current cell having movement comparable to the regular flow of liquid or gases, and ignore the number and the distribution of MUs in the neighbouring cells.

Kown et al. (2003) also used the SRWM-based mobility model and estimated the cell overload probability through considering the number of MUs in current left and right cells using the convolution sum of binary handoff probability distribution functions for the cellular network covering the highways. The CAC policy of this scheme restricts the admission of new services into the cell when the cell overload probability is less than a given predefined threshold. Naghshineh and Schwartz (1996) further improved the SRWM-based CAC policy by using the distributed information including the number of existing services in the current and neighbouring cells. The SRWM-based CAC schemes presented above are designed for a single type traffic (i.e., not for multimedia applications) and regulate the traffic loads through blocking the new services based on some statistically expected measurements of congestions. These schemes, therefore, do not reserve any resources for handoff services and cannot adapt to the changes in network conditions, as the true traffic scenario or mobility information is not considered directly.

A dynamic resource reservation scheme is presented in some schemes (e.g., Oliviera et al., 1998) where resources are reserved explicitly for each service in the neighbouring cells where any statistical or probabilistic information is not considered. The reservation of resources in this scheme is released only when either the cell goes outside the considerable neighbouring cells of the MU or the connection terminates. A huge amount of resources become wastage due to explicit reservation, as the MU will not move all of the neighbouring cells for the given time. The SRWM-based resource reservation and CAC scheme presented by Fang and Zhang (2002) uses a probabilistic approach to predict the amount of resources to be reserved in each cell using statistical measurements of handoff probability. The scheme divides the mobile users into two classes based on their speeds: the high-speed user and the low-speed user. The high-speed user has a speed more than a random walk. Any user who is stationary or moving at walking speed is considered as the low-speed user. The average cell dwelling time of a high-speed user is shorter than that of a low-speed user. So, a high-speed user is more likely to request a handoff than a low-speed user. Let $f_h(t)$ and $f_l(t)$ denote the cell dwelling time probability density function (pdf) of the high-speed and the low-speed users respectively. The handoff probability for the high-speed user is:

$$L_h(t, T) = \int_0^{T-t} f_h(\tau) d\tau \quad (1)$$

Similarly, we can obtain $L_l(t, T)$ by substituting $f_h(t)$ with $f_l(t)$. Now the total influence that all the ongoing services in cell i exerting on cell j is:

$$I_{i,j} = \sum_{k \in S} a_{i,j} L(t_k, T) \quad (2)$$

where S is the set of all currently ongoing services in cell i , $L(t_k, T)$ can be either $L_h(t, T)$ or $L_l(t, T)$, and $\alpha_{i,j}$ is the directional factor—that is, the probability that the handoff target cell is cell j when the service is being served in cell i . For a homogenous cellular network, $\alpha_{i,j} = 1/6$. In this scheme, it is assumed that the number of channels needed to be reserved is proportional to the extent of the influence (i.e., $R_{i,j} = B I_{i,j}$), where B is a tunable constant. Hence at time T , cell j needs to reserve $R_j = \sum_{i \in N_j} R_{i,j}$ channels for

the possible handoff services from its neighbouring cells, N_j . The scheme, however, assumes that the cell dwelling time is negative exponentially distributed—that is, $f_h(\tau) = \mu_h e^{-\mu_h \tau}$ and $f_l(\tau) = \mu_l e^{-\mu_l \tau}$, where $1/\mu_h$ and $1/\mu_l$ are the average cell dwelling time for high-speed and low-speed users respectively.

All the schemes reserving resources or regulating the network traffics through blocking the new services without using any mobility information cannot optimally utilize the resources due to considering the randomness in representing the behaviour of the MUs mobility. The movement of the MU is not random and can be predicted from the observed mobility information or parameters along with considering the road-networks and the geographical environments. The recent mobility support resource management schemes (e.g., Islam & Murshed, 2004; Aljadhari & Znati, 2001; Levine et al., 1997; Choi & Shin, 2002) predict the future movement of the MUs and reserve resources accordingly though forming a shadow cluster. The following sections explain the techniques of mobility prediction, shadow cluster formation, and resource reservation, along with CAC policies in detail.

FUTURE MOBILITY PREDICTION TECHNIQUES

The future mobility prediction in mobile networks implies the estimation of the probability of an MU to visit a given neighbouring cell. This probability is known as the cell visiting probability (CVP) (see Islam & Murshed, 2004; Islam, Murshed, & Dooley, 2003, 2004). Estimation of the CVP is very essential for the resource management in cellular networks, as this probability can be used to identify the neighbouring cells and compute the amount of resources that need to be reserved in those cells. The CVPs can be two types: the time-dependent CVP and the time-independent CVP. The time-dependent CVP represents the probability of an MU to visit a given neighbouring cell at a given time. The CVP changes as time proceeds. On the other hand, the time-independent CVP refers to the probability of the MU to visit a given neighbouring cell without considering any timing information. The detailed techniques for estimating the time-dependent and the time-independent CVPs are presented below.

Time-Dependent CVP Estimation Policies

Levine et al. (1997) proposed a shadow cluster scheme which uses the individual mobility information for the first time to incorporate historical databases to estimate the time-dependent CVP. This scheme profiled user mobility information at regular intervals to estimate its handoff probability density function (pdf) (also known as boundary crossing pdf) and cell residence time pdf for each cell, and cell-independent service duration pdfs for a number of service classes. The handoff probability pdf, $\gamma_{x,i}(t)$ describes the probability that an MU x , given that it initiates in cell i or enters into it through a particular side, will remain in the cell or that it will exit the cell through a specific side.

The initial handoff probability $\gamma_{x,i|w}(t)$ represents the probability that the service originates in cell i and leaves the cell through side w . 0 in place of w indicates that the MU will not leave through any side, but will remain in the current cell i . The cell residence time pdf $g_{x,i}(t)$ is the residence time of the MU x in cell i , given that the MU either initiates in cell i or enters into it through a given side and leaves the cell through a particular side. The service duration pdf $h_{x,M(x)}(t)$ refers to the distribution of service-length for the MU x while providing a service with class descriptor $M(x)$. These pdfs are used to estimate the time-dependent CVP for the MU to visit the neighbouring cells or to remain in its current cell at a certain number of future time intervals considering all possible routes.

The CVP for an MU x (initiating an active service of class descriptor $M(x)$) in the current cell I , assuming that the MU does not move out and come back, is given as:

$$P_{i,i}^*(t) = [1 - H_{x,M(x)}(t)] \left\{ \gamma_{x,i|0}(t) + \sum_{w=1}^6 [1 - G_{x,i|w}(t)] \cdot \gamma_{x,i|w}(t) \right\} \quad (3)$$

where the cells are considered hexagonal $H_{x,M(x)}(t)$, is the cumulative distribution function for $h_{x,M(x)}(t)$, and $G_{x,i|w}(t)$ is the cumulative distribution function for $g_{x,i}(t)$ in the case that the MU x initiates in cell i and leaves it through the side w . Let $q_{x,i,j}(r|v)(t)$ be the residence time pdf for x in the cells along the route $(r|v)$ from cell i to cell j , let it be calculated using the convolution of the residence time pdfs along the “route of travel” as follows:

$$q_{x,i,m}(r|v)(t) = g_{x,i|w(r|v,1)}(t) \otimes \dots \otimes g_{x,i|v(r|v,l):w(r|v,l)}(t) \otimes \dots \otimes g_{x,m|v(r|v,m):w(r|v,m)}(t) \otimes g_{x,j|v(r|v,k):w(r|v,k)}(t) \quad (4)$$

The probability of taking a particular route ($r|v$) (i.e., $R_{x,r|v}$) can be calculated by multiplying the initial handoff probability with the handoff probabilities in each cell along the route, for example:

$$R_{x,r|v} = \gamma_{x,i \rightarrow w(t|i)}(t) \prod_l \gamma_{x,l \rightarrow v(t|i)}(t) \quad (5)$$

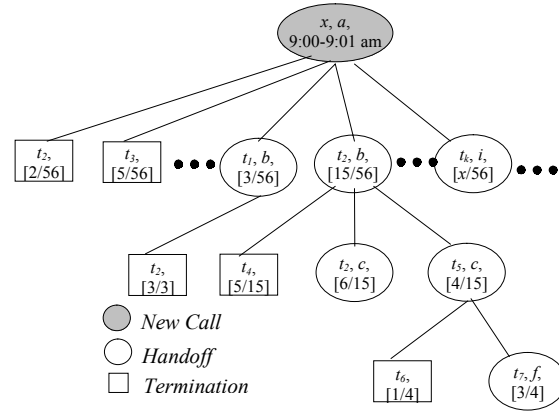
The CVP of the MU x currently in cell i to visit the cell j at time t (i.e., $P_{i,j}^x(t)$) can then be found as follows:

$$P_{i,j}^x(t) = \left[1 - H_{x,M(x)}(t) \right] \sum_{v=1}^6 \sum_r R_{x,r|v} \cdot \left(\sum_{w=1}^6 \left[Q_{x,i,j|w}(t) - Q_{x,i,j|w}(t) \right] \gamma_{x,i,j|w} \right) \quad (6)$$

where $Q_{x,i,j|w}(t)$ is the cumulative pdf of $q_{x,i,j|w}(t)$ and m is the last cell along the route ($r|v$) before the cell j .

The above-mentioned CVP estimation technique is computationally very expensive, if not impossible. It requires constant updating of a large number of pdfs for each MU from the profiled databases. Moreover, the number of possible routes from any cell to one of its neighbour could be infinite unless a finite number of shortest or the most popular paths are considered, or the exact travelling route is known from a vehicle navigational system. Yu and Leung (2002) improved the time-dependent CVP estimation technique (see Levine et al., 1997) using the Ziv-Lempel data compression algorithm (see Ziv & Lempel, 1978) on its profiled databases organised as a collection of tries, one for each cell. The scheme represents the lifetime of a service as an event sequence such as $(N(x,i,t_0), H(x,j_1,t_{k_1}), H(x,j_2,t_{k_2}), \dots, H(x,j_n,t_{k_n}), \dots, E(x,t_k))$, where $N(x,i,t_0)$ represents that the MU x is originated in cell i at time t_0 , $H(x,j_n,t_{k_n})$ represents that the MU x will handoff to cell j_n after a time period of t_{kn} , and

Figure 3. A mobility trie used for mobility prediction



$E(x,t_k)$ is the termination event of x after a time period of t_k . The scheme observed the event-sequences in the past to create a mobility trie for each MU originated in a given cell. Each node of the trie except the root preserves the relevant statistical information that can be used to predict the probabilities of all possible events and compute the expected time-dependent CVPs.

For an instance, assume that a mobility trie for an MU x originated in cell a at the time interval t_0 (say, 9:00-9:01 a.m.), as shown in Figure 3. Also assume that three cells a , b , and c are adjacent to each other and the cell f is adjacent to cell c . The information (t_k, p) in the leaf node of the trie implies that the MU x will terminate after a time period of t_k with probability p , and (t_k, j, p) in the interior node represents that the MU x will be handed off to cell j with the probability p after a time period of t_k . Table 1 shows the different possible paths, along with the probability of taking those paths by the MU x to move into the cell b from the cell a .

The time-dependent CVPs of the MU x (currently in cell a) to visit the cell b at the time

Table 1. The probability estimation for an MU to follow a given path

Path Name	Sequence of Events	Probability
Path 1	$(N(x, a, t_0), H(x, b, t_1), E(x, t_2))$	$\frac{3}{56} \times \frac{3}{3} = \frac{3}{56}$
Path 2	$(N(x, a, t_0), H(x, b, t_2), E(x, t_4))$	$\frac{15}{56} \times \frac{5}{15} = \frac{5}{56}$
Path 3	$(N(x, a, t_0), H(x, b, t_2), H(x, c, t_2), \dots)$	$\frac{15}{56} \times \frac{6}{15} = \frac{6}{56}$
Path 4	$(N(x, a, t_0), H(x, b, t_2), H(x, c, t_5), \dots)$	$\frac{15}{56} \times \frac{4}{15} = \frac{4}{56}$

period t ($t = t_1, t_2, t_3, t_4$, and t_5) (i.e., $P_{a,b}^x(t)$) based on the paths shown in Table 1 are given in 7.

$$\begin{aligned}
 P_{a,b}^x(t_1) &= \frac{3}{56} \\
 P_{a,b}^x(t_2) &= \frac{3}{56} + \frac{5}{56} + \frac{6}{56} + \frac{4}{56} = \frac{18}{56} \\
 P_{a,b}^x(t_3) &= \frac{5}{56} + \frac{4}{56} = \frac{9}{56} \\
 P_{a,b}^x(t_4) &= \frac{5}{56} + \frac{4}{56} = \frac{9}{56}, \text{ and} \\
 P_{a,b}^x(t_5) &= \frac{4}{56}
 \end{aligned} \tag{7}$$

Although the Zip-Lempel algorithm inherently considers only the routes taken so far by the MU from its current cell to the destination, still the number could be high enough to make the estimation process computationally infeasible. The time-dependent CVP estimation techniques presented above, however, profile the user movements and estimate CVP for each MU in each cell for every time period. These policies are, therefore, extremely difficult and time consuming, if not impossible. So, most of the resource management schemes estimate the time-independent CVP along with the time-window, for timing information, in which the MU is highly expected to visit the cell.

Time-Independent CVP Estimation Policies

Estimation of time-independent CVP is a comparatively more feasible, reliable, and efficient way for predicting the future movement of an MU. Many schemes (e.g., Islam & Murshed, 2004; Aljadhari & Znati, 2001; Choi & Shin, 2002) have been presented in literature to measure the time-independent CVP from the historical databases/movement patterns and the key mobility parameters (e.g., position, direction, and speed). The major time-independent CVP estimation techniques using the different types of available information to estimate the future mobility of the MU are presented below.

Time-Independent CVP Estimation from the Historical Databases/ Movement Patterns

Using the historical databases and/or movement patterns is a conventional technique for estimating the time-independent CVP. Tabbane (1995) used user-specific mobility profiling databases for dynamic location tracking based on the quasi-deterministic mobility behaviour represented as a set of movement patterns. This method was further pursued by Liu and Maguire (1995) to model mobility behaviour as the repetitions of some elementary movement pat-

terns. Based on these movement patterns, a pattern-matching/recognition-based mobile motion prediction (MMP) algorithm is proposed to estimate the location of the MU. The MMP algorithm is, however, highly sensitive to so-called random movements, which are defined as the movement that cannot be classified by the simple mobility patterns (see Liu, Bahl, & Chlamtac, 1998). Moreover, the use of the movement patterns leads to huge databases and does not include the MU's current movement trends.

Choi and Shin (2002) demonstrated the influences of cell topography and road networks in handoff probability estimation with the assumption that the mobility behaviour of an MU will be probabilistically similar to that of the MUs which came from the same previous cell and are now residing in the current cell. Hence, instead of profiling mobility history for each MU separately, the scheme proposed to profile up to a predefined number of handoff history—that is, quadruplet consisting of the departing time from the current cell, the index of the previous cell the MU resided in before entering the current cell, the index of the cell the MU entered after departing the current cell, and the sojourn time (i.e., the current cell dwelling time). The CVP of an MU to visit a neighbouring cell is then estimated from the aggregated history of handoff observed as:

$$\frac{\text{Number of past services observed in the current and eventually visited the given cell}}{\text{Number of services in the curr}}$$

(8)

This feasible scheme uses very limited information to estimate CVP, and ignores the mobility direction and speed variation of the current MU as well as the direct effects of cell topography or road networks.

Time-Independent CVP Estimation from the Directional Information Only

Aljadhah and Znati (2001) used the direction information for estimating the time-independent CVP for an MU x in any neighbouring cell j . To overcome the “large” effects in estimated CVP due to small deviations in the MU's direction, the scheme predicts a future mobility direction based on the statistics of the movement history. Let D_0 be the current direction of the MU, D_t represents the observed direction of the MU at time t , and $\tilde{D}_t$ is the estimated direction at time t , then the predicted direction $\tilde{D}_{t+1}$ at $(t+1)$ is obtained as follows:

$$\tilde{D}_{t+1} = (1 - \alpha)\tilde{D}_t + \alpha D_t \quad (9)$$

where α is the smoothing factor of the first-order autoregressive filter used in (9). In order to quickly track down abrupt changes in the actual direction of the MU, α is computed as:

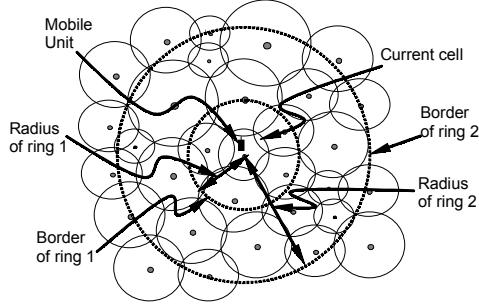
$$\alpha = c \frac{E_t^2}{\sigma_{t+1}} \quad (10)$$

where $0 < c < 1$, $E_t = D_t - \tilde{D}_t$ is the prediction error, and σ_t is the average of the past square prediction errors at time t . To make the prediction process more adaptive, σ_t can be expressed as:

$$\sigma_{t+1} = cE_t^2 + (1 - c)\sigma_t \quad (11)$$

The use of first-order autoregressive filter to predict future direction along with the estimation technique of its smoothing factor guarantees that the predicted mobility direction is not affected by small deviations in observed direction.

Figure 4. Definition of rings in heterogeneous cellular networks



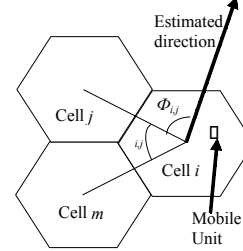
To estimate the CVP, the scheme divides the neighbouring cells into different imaginary rings, as shown in Figure 4. The radius of the first ring with respect to cell i , $R_{i,1}$ is equal to the distance from the centre of cell i to the centre of the adjacent neighbouring cells whose centre is furthest away. The radius of the n th ring $R_{i,n}$ is equal to n times the radius of $R_{i,1}$. Any cell that has its centre within the boundary of a ring is considered to be in that ring.

According to this scheme, the CVP of an MU x to visit a neighbouring cell j is estimated from the current cell i , the immediate previous cell m , and the estimated mobility direction $\tilde{D}_i$. Let the centre of cell j be situated at an angle $\phi_{i,j}$ from $\tilde{D}_i$, and $\theta_{i,j}$ from the line joining the centre of cell i and the centre of cell m , as shown in Figure 5. A directionality estimate for cell j can be measured as:

$$D_{i,j}^x = \begin{cases} \frac{\theta_{i,j}}{\phi_{i,j}}, & \phi_{i,j} > 0 \\ \theta_{i,j}, & \phi_{i,j} = 0 \end{cases} \quad (12)$$

The CVP $P_{i,j}^x$ can be found through normalizing the directionality estimates over the ring in which the cell j resides:

Figure 5. Parameters used to calculate CVP from directional information



$$P_{i,j}^x = \frac{D_{i,j}^x}{\sum_k D_{i,k}^x} \quad (13)$$

where k is the cell at the same ring of the cell j with respect to cell i . Note that a cell k is said to be at ring L with respect to cell i if it is located at a ring L cells away from the cell i .

The CVP estimation technique presented in this scheme is, however, restrictive for the following reasons. First, the intra-cell mobility is not considered, as the directionality is measured from the centres of cells, irrespective of the position of the MU inside the current cell. Second, due to considering immediate previous cell centre in the directionality estimation, no CVP can be calculated unless an inter-cell movement path is established. Third, only direction is used while ignoring the other important key mobility parameters such as speed and travelling distance. As a result, CVPs of all the neighbouring cells whose centres lie along the same line as that of the current cell will be identical irrespective of their distance from the MU.

Time-Independent CVP Estimation from the Key Mobility Parameters: Direction, Position, and Speed Information

Islam et al. (2003, 2004) incorporated the distance information and the speed of the MU along the with the predicted direction of mobility (9) to compute the time-independent CVP. Instead of using the aerial distance, the schemes propose to estimate the expected travelling distance (ETD) to include the effects of road-networks and geographical environments. In the schemes, the travelling distance of the MU x currently in cell i to visit the cell j is measured as:

$$S_{i,j}(x) = (T_j^{arr}(x) - T_i^{act}(x))v_x \quad (14)$$

where $T_j^{arr}(x)$ is the arrival time of the MU x in cell j , $T_i^{act}(x)$ is the first activation time of x in cell i , and v_x is the average speed of x . Let $\tilde{S}_{i,j}^k$ be the ETD, in state k , travelled by x while moving from cell i to cell j . The ETD is updated using the exponential averaging function (EAF) as shown in 15 every time x arrives in cell j after being activated in cell i as follows:

$$\tilde{S}_{i,j}^{k+1} = (1 - \omega_{i,j}^k) \tilde{S}_{i,j}^k + \omega_{i,j}^k S_{i,j}(x) \quad (15)$$

where $\tilde{S}_{i,j}^0 = s_{i,j}$ is the aerial distance from cell i to cell j and $\omega_{i,j}^k$ is the weight of EAF, also known as the distance smoothing factor (DSF), in state k . The DSF is a knowledge-based learning parameter and determines the sharing of knowledge for the past travelling distances. If we consider $\omega_{i,j}^k = 1/n_{i,j}^k$, where $n_{i,j}^k$ is the total number of observations up to the state k , the EAF shown in 15 becomes the linear averaging function. In order to estimate the ETD more accurately, the DSF is computed dynamically from the normalized distance error $\epsilon_{i,j}^k = (\tilde{S}_{i,j}^k - \tilde{S}_{i,j}^{k-1})/\tilde{S}_{i,j}^k$ and the exponential av-

erage of the previous normalized distance squared errors $\xi_{i,j}^k$, as shown in 16:

$$\omega_{i,j}^k = c \frac{(\epsilon_{i,j}^k)^2}{\xi_{i,j}^k + 1} \quad (16)$$

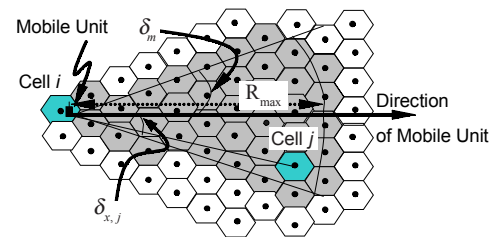
where $0 < c < 1$ is called the sensitivity parameter to handle the abrupt changes in the communication environments. The normalized errors in 16 make the ETD estimation technique more general irrespective of traffic patterns in a specific road. Moreover, to make the ETD estimation technique flexible to the environmental changes, $\xi_{i,j}^k$ is updated again using EAF while considering c as a smoothing factor in 17:

$$\xi_{i,j}^k = c(\epsilon_{i,j}^k)^2 + (1-c)\xi_{i,j}^{k-1}, \quad \xi_{i,j}^0 = 0. \quad (17)$$

During the transition period between the off-peak and the busy hours or in increase of changing traffic behaviours, the value of c should be higher to give more priority to the current traffic nature rather than historical traffic information; conversely, if the traffic shows its regular behaviour, the value of c should be lower to consider long-time historical feedback.

The schemes presented by Islam et al. (2003, 2004) compute time-independent CVP though

Figure 6. Graphical definition of $\delta_{x,j}$ and formation of the shadow cluster (the shaded area)



assuming some basic properties of the CVP distribution as shown in (18), (19), and (20), where $\delta_{x,j}$ is the directional angle between the predicted direction of x and a line joining the centre of cell j with the position of x , as shown in Figure 6.

P-I: $P_{i,j_1}^x \geq P_{i,j_2}^x$ if and only if $\delta_{x,j_1} \leq \delta_{x,j_2}$ —that is, the CVP is the maximum in those cells, lies in the direction of mobility and decreases as $\delta_{x,j}$ increases, and equals almost zero when the directional angle $\geq \pi/2$. This property can be represented as:

$$P_{i,j}^x \propto \cos \delta_{x,j}, \quad \delta_{x,j} \leq \pi/2 \quad (18)$$

assuming for the constant velocity and ETD, the CVP of an MU x in a cell j is directly proportional to the cosine of the directional angle of cell j measured with respect to the direction of mobility.

P-II: For a specific directional angle and ETD, $P_{i,j}^x$ is proportional to v_x :

$$P_{i,j}^x \propto v_x \quad (19)$$

P-III: For a specific directional angle, $P_{i,j}^x$ is inversely proportional to *average time* to reach the centre of cell j :

$$P_{i,j}^x \propto \tilde{S}_{i,j} / v_x \quad (20)$$

where $\tilde{S}_{i,j}$ is the ETD at the current state.

Through combining 18, 19, and 20, we can get the relationship of $P_{i,j}^x$ with v_x , $\delta_{x,j}$, and $\tilde{S}_{i,j}$ as shown in 21:

$$P_{i,j}^x \propto \frac{v_x^2 \cos \delta_{x,j}}{\tilde{S}_{i,j}}, \quad \text{if } \delta_{x,j} \leq \pi/2 \quad (21)$$

The proportionality relation in (21) can be replaced to an equal relationship by including a proportionality constant K as follows:

$$P_{i,j}^x = \begin{cases} K \frac{v_x^2 \cos \delta_{x,j}}{\tilde{S}_{i,j}}, & \text{if } \delta_{x,j} \leq \pi/2 \\ 0 & \text{otherwise} \end{cases} \quad (22)$$

where the value of K can be found through normalising $P_{i,j}^x$ over the imaginary ring (as shown in Figure 4) which includes cell j . So, after finding the value of K and placing it in (22), the CVP can be found as:

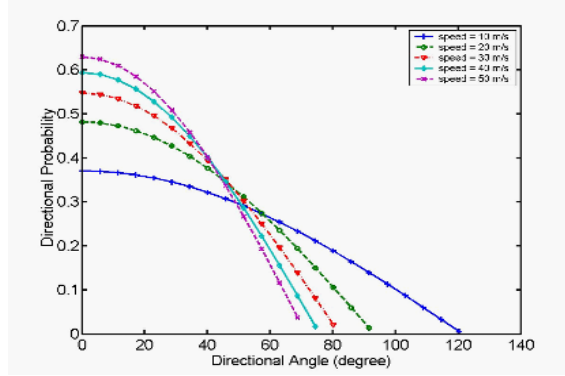
$$P_{i,j}^x = \begin{cases} \frac{\left(\frac{v_x^2 \cos \delta_{x,j}}{\tilde{S}_{i,j}} \right)}{\sum_{k \in R_{i,n} | j \in R_{i,n}} \left(\frac{v_x^2 \cos \delta_{x,k}}{\tilde{S}_{i,k}} \right)}, & \text{if } \delta_{x,j} \leq \pi/2 \\ 0 & \text{otherwise} \end{cases} \quad (23)$$

where the condition on k implies that cell k is at same ring of the cell j with respect to cell i .

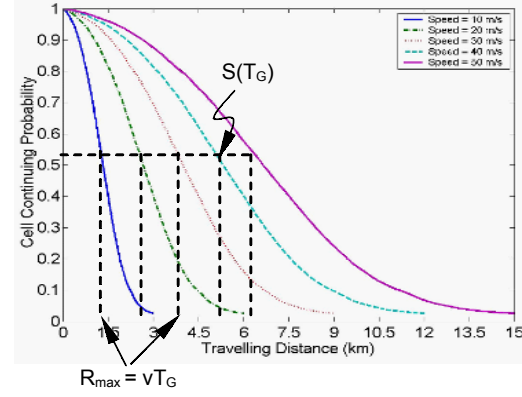
The schemes (Islam et al., 2003, 2004) are further extended (see Islam & Murshed, 2004; Islam et al., 2004) through estimating the CVP by computing and combining the directional probabilities (DPs) and call continuing probabilities (CCPs). These schemes, however, consider the major influences of the mobility parameters on the MU's future movement. The schemes estimate DPs as shown in (24) based on the assumption that: (1) the CVP is always the highest in the direction of mobility and decreases as $\delta_{x,j}$ increases—that is, $P_{i,j_1}^x \geq P_{i,j_2}^x$ if and only if $\delta_{x,j_1} \leq \delta_{x,j_2}$; and (2) the rate of decrease in probability, with the increase in $\delta_{x,j}$, is proportional to the speed, where the rate of decrease is the largest (smallest) for the highest (slowest) speed considered.

$$DP_{i,j}^x = \begin{cases} \frac{\Phi}{2} \cos \Phi \delta_{x,j}, & \text{if } \delta_{x,j} \leq \frac{\pi}{2\Phi} \text{ and } \frac{1}{2} \leq \Phi \leq 2; \\ 0, & \text{otherwise} \end{cases} \quad (24)$$

Figure 7. The characteristics of directional probabilities and call continuing probabilities



a. Directional probabilities at different speed.



b. Call continuing probabilities at different speed.

where $\Phi = \log v_x / \log \sqrt{v_{\max}}$ is the *speed sensitivity factor* and $v_{\max}$ is the maximum speed permitted under the scheme. Note that the use of cosine function validates assumption 1 and the logarithmic speed sensitivity factor Φ validates assumption 2. The angular span of DP decreases with the increasing speed. Now, $\Phi \geq \frac{1}{2}$ as the maximum angular span is 180° and $\Phi \leq 2$ as the probability is bounded by unity.

The CCP is estimated using the survival function of the Gaussian distribution representing the call termination probability. Let $\bar{S}_{i,j}$ be the mean travel distance and $\text{var}(S_{i,j})$ be the travel distance variance calculated from the past observations of all the active MUs travelled from cell i to cell j using exponential averaging functions. Let $f(t)$ and $F(t)$ be the pdf and cumulative pdf of the call duration time with mean TG. The $CCP_{i,j}^x$ is measured from the survivality function $S(t) = 1 - F(t)$ as follows:

$$CCP_{i,j}^x = S(\bar{S}_{i,j} / v_x) \quad (25)$$

The DP and CCP are correlated, as both the probabilities are calculated using the speed of

the MU. Therefore, these two probabilities can be combined to calculate CVP, $P_{i,j}^x$, as follows:

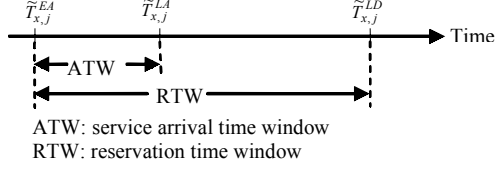
$$P_{i,j}^x = (DP_{i,j}^x)^\mu (CCP_{i,j}^x)^{1-\mu} \quad (26)$$

where $0 < \mu < 1$ is the tuning parameter to adjust the dependencies between the DP and the CCP. The value of μ close to 1.0 implies that the DP is dominating over the CCP and close to 0 implies that the CCP has more influence in estimated CVP. So, to get a good contribution from both the DP and the CCP, a value $\mu = 0.5$ is the best compromise.

RESERVATION TIME WINDOW ESTIMATION

As the time-independent CVP (e.g., Islam & Murshed, 2004; Aljadhahi & Znati, 2001) cannot provide any timing information, it is obvious to estimate a reservation time window (RTW) for which the resources are reserved in the corresponding cell for the given MU. For estimating the RTW, the scheme proposed by Aljadhahi and Znati (2001) estimates the expected earliest

Figure 8. Resource leasing and reservation intervals within a cell



arrival time ($\tilde{T}_{x,j}^{EA}$), the expected latest arrival time ($\tilde{T}_{x,j}^{LA}$), and the expected latest departure time ($\tilde{T}_{x,j}^{LD}$) for the MU x in cell j . The time interval $[\tilde{T}_{x,j}^{EA}, \tilde{T}_{x,j}^{LD}]$ is the maximum expected residence time of the MU within cell j known as the RTW or resource reservation interval. The resources are reserved in the cell for the entire period of RTW. The time interval $[\tilde{T}_{x,j}^{EA}, \tilde{T}_{x,j}^{LA}]$ is known as the service arrival time window (ATW) or resource leasing interval, as all the reserved resources are released if the MU does not arrive within this time period. This is required to prevent the MUs from holding the resources unnecessarily. Figure 8 shows the RTW and the ATW for a given MU in a cell.

To get $\tilde{T}_{x,j}^{EA}$, $\tilde{T}_{x,j}^{LA}$, and $\tilde{T}_{x,j}^{LD}$ for deriving the RTW and the ATW, the scheme considers all possible paths from current cell i to each of the neighbouring cells under consideration. Although there may be many paths to go from any cell to one of its neighbouring cells, the scheme simplifies the RTW and the ATW estimation process based on the assumption that there is only one path between two adjacent cells. The neighbouring cells, including the current cell, are represented as a directed graph $G = (V, E)$ where a vertex $v_j \in V$ represents a cell j and an edge $v(v_i, v_j) \in E$ represents the direct path from cell i to its adjacent neighbour cell j , if any. This method is highly complex and time consuming, as there are many possible paths that an MU may follow to reach a neighbouring cell (see Aljadhari & Znati, 2001). An acceptable and

simplified solution can be found using the k-shortest paths algorithm (see Dreyfus, 1969) to obtain a finite number of most likely paths to be followed by the MU.

After selecting a finite number of paths from the current cell i to a neighbouring cell j , the path residence time is estimated for each of the paths as the sum of the residence time in each cell along the path. The expected earliest arrival time $\tilde{T}_{x,j}^{EA}$ and the expected latest arrival time $\tilde{T}_{x,j}^{LA}$ are then estimated as the smallest and the largest path residence time as shown in (27).

$$\tilde{T}_{x,j}^{EA} = \sum_{k \in \Pi_s} \frac{d(m, k, n)}{\bar{S}_{\max}(k)} \quad \text{and} \quad \tilde{T}_{x,j}^{LA} = \sum_{k \in \Pi_l} \frac{d(m, k, n)}{\bar{S}_{\min}(k)} \quad (27)$$

where Π_s and Π_l are the paths with the smallest and the largest path residence times, respectively. The average maximum speed $\bar{S}_{\max}(k)$ and the average minimum speed $\bar{S}_{\min}(k)$ in cell k are found as the maximum and the minimum speeds of the MUs observed in the past. The value of $d(m, k, n)$ —that is, the mean distance within cell k given that the cells m , k , and n are three consecutive cells—depends on whether the cell k is the originating, the intermediate, or the last cell in the path and can be found as

$$d(m, k, n) = \begin{cases} d_o(k, n), & \text{if } k \text{ is the originating cell, } k = i \\ d_l(m, k, n), & \text{if } k \text{ is an intermediate cell} \\ d_{lp}(m, k, n), & \text{if } m = n \\ d_l(m, k), & \text{if } k \text{ is the last cell, } k = j \end{cases} \quad (28)$$

The value of $d_o(k, n)$ is approximated as the mean distance (i.e., $8R/3\pi$), based on the assumption that the MUs are evenly spread over a cell area of radius R , travel along a constant direction within the cell, and exit from the cell through any point along the border of cell n . To find the value of $d_l(m, k, n)$, it is assumed that the MU enters from cell m to cell k at any point

along the arc AB, travels in cell k along a constant direction, and then exits the cell k through any point along the arc CD, as shown in Figure 9. Let the arc AB be defined by the angles β_1 and β_2 , and the arc CD be defined by the angles θ_1 and θ_2 . The mean distance of the MU inside the cell k (i.e., $d_l(m, k, n)$) is estimated as shown in 29:

$$d_l(m, k, n) = \begin{cases} \frac{8R}{(\theta_2 - \theta_1)(\beta_2 - \beta_1)} \left[\sin\left(\frac{\beta_2 - \theta_2}{2}\right) - \sin\left(\frac{\beta_2 - \theta_1}{2}\right) \right], & \text{for } \beta_1 \geq \theta_2 \\ \frac{8R}{(\theta_2 - \theta_1)(\beta_2 - \beta_1)} \left[\sin\left(\frac{\theta_1 - \beta_2}{2}\right) - \sin\left(\frac{\theta_1 - \beta_1}{2}\right) \right], & \text{for } \beta_2 \leq \theta_1 \end{cases} \quad (29)$$

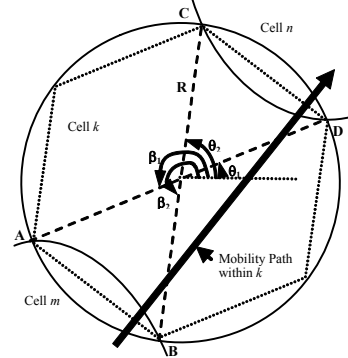
The mean distance in the last cell in the path is derived as $d_L(m, k) = \max d(m, k, q)$ for $\forall q$ adjacent to k, $q \neq m$ (see Aljadhari & Znati, 2001). The distance in cell k for $m = n$ is found as $d_{LP}(m, k, n) = 2.d_O(k, n) = 16R/3\pi$ considering the path as a loop inside the cell k.

The expected latest departure time (i.e., $\tilde{T}_{x,j}^{LD}$) can be found by adding the maximum cell residence at the last cell k of the path with $\tilde{T}_{x,j}^{LA}$ as follows:

$$\tilde{T}_{x,j}^{LD} = \tilde{T}_{x,j}^{LA} + \frac{d_L(m, k)}{\bar{S}_{\min}(k)} \quad (30)$$

As the values of $\tilde{T}_{x,j}^{EA}$, $\tilde{T}_{x,j}^{LA}$, and $\tilde{T}_{x,j}^{LD}$ are estimated, the RTW and the ATW for the MU x in cell j is found from its definition presented above. However, the techniques used for estimating these values are very simplified and the assumptions used are not valid for simulating the mobility of the MU in real-world wireless networks, as there is no guarantee that an MU will always traverse along one of the finite number of shortest cell trajectories that are in constant direction within the cell. Moreover, the average minimum speed $\bar{S}_{\min}(k)$ and the average maximum speed $\bar{S}_{\max}(k)$ profiling of all cells tend to be similar, especially for larger cells.

Figure 9. Distance in an intermediate cell



Through incorporating both the road networks and the traffic flow conditions, an improved technique (see Islam & Murshed, 2004; Islam et al., 2004) is presented for estimating the RTW and the ATW from the ETD as shown in (15) and its standard deviation $\sigma_{i,j}^k$ at state k which can be found using the mathematical definition show as follows:

$$(\sigma_{i,j}^{k+1})^2 = (1 - \omega_{i,j}^k) (\sigma_{i,j}^k)^2 + \omega_{i,j}^k (S_{i,j}(x) - \tilde{S}_{i,j}^{k+1})^2 \quad (31)$$

assuming $\sigma_{i,j}^0 = 0$. Once the ETD and its standard deviation are known, a distance window can be measured by calculating the maximum ETD $\tilde{S}_{i,j}^{\max}$ and the minimum ETD $\tilde{S}_{i,j}^{\min}$ between the cell i and the cell j using the normal distribution curve, with the mean and the standard deviation as $\tilde{S}_{i,j}^x$ and $\sigma_{i,j}^k$ (the current working value of $\sigma_{i,j}^k$) respectively for a specific confidence level β , as shown in Figure 10. From the definition of confidence level in normal distribution, it can be concluded that as β increases, the confidence level of the MU to arrive into the cell increases, but the resources are reserved for a longer time, which ultimately increases the wastage of it. On the other hand, the smaller value of β decreases the RTW and increases the resource utilization, but the un-

certainty of the MU to arrive into the cell within the estimated RTW increases, which ultimately increases the CDP. For a given β , the expected earliest arrival time $\tilde{T}_{x,j}^{EA}$ and the expected latest departure time $\tilde{T}_{x,j}^{LA}$ can be found as the minimum and the maximum ends of the ATW—that is, dividing $\tilde{S}_{i,j}^{\max}$ and $\tilde{S}_{i,j}^{\min}$ by v_x respectively as shown in (32):

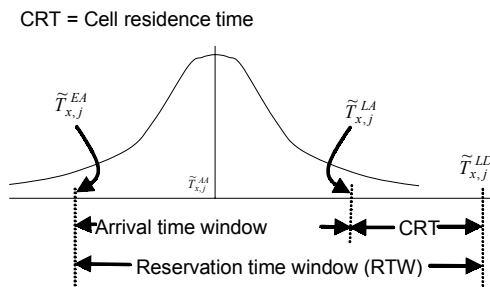
$$\tilde{T}_{x,j}^{EA} = \frac{\tilde{S}_{i,j}^{\min}}{v_x} \quad \text{and} \quad \tilde{T}_{x,j}^{LA} = \frac{\tilde{S}_{i,j}^{\max}}{v_x} \quad (32)$$

The expected earliest arrival time $\tilde{T}_{x,j}^{EA}$ defines the lower end of the RTW; whereas the higher end (i.e., $\tilde{T}_{x,j}^{LD}$) can be found by adding the maximum expected cell visiting time (CRT) $\tilde{T}_{x,j}^{MCR}$ with $\tilde{T}_{x,j}^{LA}$ as shown in (33):

$$\tilde{T}_{x,j}^{LD} = \tilde{T}_{x,j}^{LA} + \tilde{T}_{x,j}^{MCR} \quad (33)$$

The maximum expected cell visiting time $\tilde{T}_{x,j}^{MCR}$ can be estimated more accurately for each cell through considering the historical data of traffic patterns inside the cell. Figure 10 shows that the RTW equals the sum of the ATW and the CRT. The graph also shows that the arrival probability of the MU x into the cell j is the highest at the average expected arrival time $\tilde{T}_{x,j}^{AA}$ and reduces at both sides of it following the nature of a normal curve.

Figure 10. Graphical definition of RTW



FORMATION OF SHADOW CLUSTERS

Shadow cluster is a collection of neighbouring cells, each of which is characterised by the condition of having the CVP above a certain predefined threshold. For a given active MU x , the resources are reserved only in the neighbouring cells under its shadow cluster. As the active MU moves to the neighbouring cells, the shadow cluster also moves. The fundamental idea of forming a shadow cluster is that every active MU (i.e., having connection with the networks) exerts an influence in the vicinity of the current cell and along the direction of mobility, if considered. The shape and size of the shadow cluster varies in different schemes due to using different information in estimating CVP. Due to using the statistical values of the user's profiled history for estimating CVP, the shadow cluster formed in the scheme proposed by Levine et al. (1997) is stronger near the active MU and fades away depending on the distance between the MU and the cell. Zhou, Chen, and Gao (2001), however, approximated the shadow cluster through giving a specific elliptical shape, where the position of the MU is represented by one of the focus and the direction of mobility is represented by the major axis. The shadow cluster formed according to these schemes (Levine et al., 1997; Zhou et al., 2001) waste huge resources through reserving in all directions while the MU is moving in a particular direction only.

The parametric mobility support schemes (e.g., Islam & Murshed, 2004; Aljadhari & Znati, 2001), however, form the conical shape shadow cluster (also known as the forward span), as shown in Figure 6, where the maximum angular span δ_m is determined from the DP and the maximum radial span R_m is determined from the call duration information or CCP for the given thresholds. In the scheme

proposed by Aljadhari and Znati (2001), as the DP and the CVP are the same in a given cell for a given MU, the angular span of the shadow cluster is defined from the estimated CVP as follows. The angular span δ_m is the maximum directional angle (i.e., $\delta_m = \max\{\delta_{x,j}\}$) for a neighbouring cell j such that $P_{i,j}^x \geq \alpha_{TH}$, where α_{TH} is a system defined threshold on the likelihood that the cell will be visited. The maximum radial span R_m of the shadow cluster is the number of rings, as shown in Figure 4, defined as proportional to the guarantee period T_G (i.e., $R_m \propto T_G$). However, instead of using R_m , the scheme uses an adaptive radial span of $R_k \leq R_m$ rings while forming a shadow cluster to support the users moving along the predicted direction. When the user deviates from the estimated direction, R_k is decreased by an amount proportional to the degree of deviation. The deviation Dev_t is measured as:

$$Dev_{t+1} = \beta.Dev_t + (1 - \beta)|\tilde{D}_t - D_t| \quad (34)$$

where $0 < \beta < 1$ and Dev_0 equal zero. The adaptive radial span at time t is defined as

$$R_k = \left\lfloor \left(1 - \frac{Dev_t}{\pi}\right)^2 . R_m \right\rfloor \quad (35)$$

The shadow cluster or forward span in this scheme includes all cells j within the angular span of δ_m and the radial span of R_k . This estimated shadow cluster is still erroneous as the estimation technique does not consider any speed information—an important factor to be considered. Due to considering the speed-independent DP for the angular span and the guaranteed period TG for the radial span, the shape of the shadow cluster and the number of cells within it remains same, regardless of whether the MU is moving at a very low or high speed. Therefore, if the size of the shadow cluster is

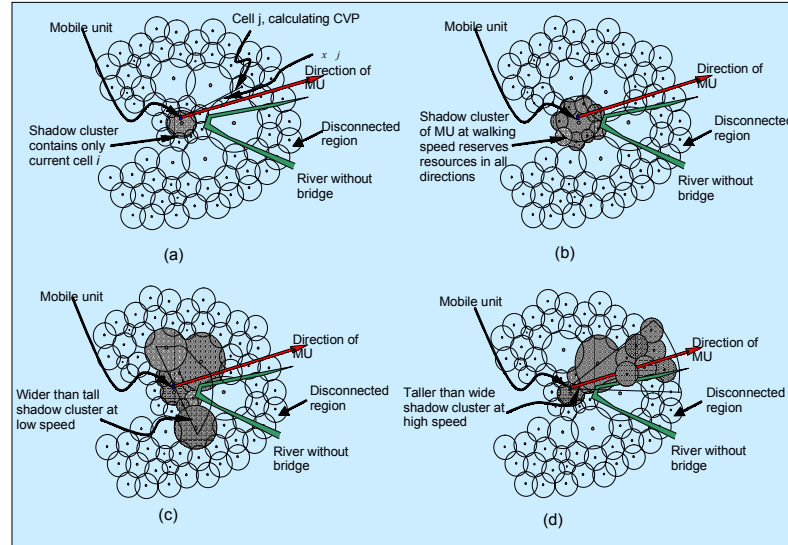
optimized for very low speed units, the CDP for high-speed MUs will increase significantly. Conversely, if the size of the shadow cluster is optimized for high-speed units, the CBP and the channel utilization will degrade significantly for low-speed units due to excessive reservation. Moreover, the scheme unfairly punishes the users who are not moving along the predicted direction of mobility, as it is the system's fault for not being able to predict it accurately.

Islam and Murshed (2004) proposed a further improved technique of forming the conical shaped shadow cluster through estimating the angular span δ_m from and the radial span R_m from $DP_{i,j}^x \geq \tau_{ap}$, as shown in 36, where τ_{ap} is a predefined probability threshold:

$$\begin{aligned} \delta_m &= \arg \min_{\forall \delta_{x,j}} (DP_{i,j}^x | DP_{i,j}^x \geq \tau_{ap}) \\ \text{and } R_m &= \max (\tilde{S}_{i,j}^x | \tilde{S}_{i,j}^x \leq v_x T_G) \end{aligned} \quad (36)$$

As both the DP and the CCP are speed dependent, the shadow cluster formed in this scheme highly depends on speed information—that is, the shadow cluster is taller-than-wide for the high-speed units and wider-than-tall for the low-speed units. Figure 11 shows the shapes and sizes of the shadow clusters at different speeds. If the MU is stationary, it contains only the current cell. When the MU is moving at a walking speed, the shadow cluster contains the current and the adjacent cells only. A wider angular span and shorter radial span shadow cluster is formed for low-speed MUs as the unit has higher probability to change its direction and cover a shorter distance. The MU moving at higher speed covers longer distance along the direction of mobility and has a small tendency to change its direction. In such a case, a taller-than-wide shadow cluster is formed. It is interesting to note that, as distance travelled is considered instead of just a straight line distance (see Islam & Murshed, 2004), an adja-

Figure 11. The shape of size of shadow cluster at different speeds. The shadow cluster (a) contains just the current cell when the MU is stationary, (b) contains the current cell and the contact neighbouring cell when the MU is moving at walking speed, (c) is wider-than-tall at low speed, and (d) is taller-than-wide at high speed. Note that the disconnected regions are not included in the shadow cluster in (c) and (d) due to considering the ETD in CVP estimation.



cent cell, which is disconnected from the road network, will not be considered in the shadow cluster, while a cell connected with high-speed freeway is included in the shadow cluster even when it is much further away than many others cells that are not included. So, the shadow cluster formed by the boundary conditions given in (53) is adaptive with the speed, direction, and distance, as well as with the geographical conditions and, hence, saves resources from being wasted due to unnecessary reservations.

RESOURCE RESERVATION POLICIES

For efficient resource management, it is unavoidable to estimate the amount of resources needed to be reserved in each cell to support the handoff services. The GC resource reser-

vation scheme, as discussed earlier, reserves a fixed portion of the cell capacity (i.e., a fixed number of channels) for the handoff services without considering the mobility information or the number of services in the current and the neighbouring cells. Many recent schemes (e.g., Islam & Murshed, 2004; Aljadhari & Znati, 2001) use mobility information to estimate the amount of resources needed to be reserved in each cell for a given MU in such a way that the minimum amount of resources is reserved based on the estimated CVP. To accommodate the multimedia services, instead of using a channel for each service, these schemes divide the cell capacity into smaller units, known as bandwidth units (BUs), which is the minimum quota of bandwidth resources. A service may use more than one BU depending on the required transmission rate. Let $B_x^{\max}$ be the desired BUs for a service x and $X_i \subset X$ be the set of all active MUs

residing in cell i , where X includes all the MUs either residing in cell i or reserving resources in it. The resource reservation scheme presented by Aljadhari and Znati (2001) reserved a fixed portion (say, $\gamma\%$) of $B_x^{\max}$ for a given MU x in each cell of the corresponding shadow cluster. According to this scheme, the total amount of resources reserved in a cell i (i.e., C_i^r) is given by:

$$C_i^r = \sum_{y \in X \text{ and } y \notin X_i} \gamma \cdot B_y^{\max} \quad (37)$$

Reserving a fixed portion of desired bandwidth for each MU, however, ignores the CVP of the MU to that cell. As the scheme estimates the CVP from the directional information only, ignoring the distance of the corresponding cell from the MU together with the MU's average speed, there is no other alternative but to suggest a QoS guarantee service model reserving the fixed portion of the MU's desired resources. Bandwidth utilization along with the CBP and the CDP can be improved by adopting the variable proportion desired bandwidth reservation policy depending on the CVP, by incorporating information concerning the relevant key mobility parameters. The schemes (e.g., Levine et al., 1997; Yu & Leung, 2002) estimate the amount of resources reserved for a given MU in a cell as the CVP times the desired bandwidth for each time-slot t . So, the total bandwidth reserved in a given cell i is:

$$C_i^r(t) = \sum_{y \in X \text{ and } y \notin X_i} P_{i,j}^y(t) \cdot B_y^{\max} \quad (38)$$

Estimating the amount of resources reserved in each time slot for each cell is complex and time consuming, as is the estimation process of the time-dependent CVP. The scheme proposed by Islam and Murshed (2004) has extended the resource reservation policy through estimating the amount of resources reserved

using the time-independent CVP. Based on this scheme, total amount of resources reserved in cell i is given by (39):

$$C_i^r = \sum_{y \in X \text{ and } y \notin X_i} P_{i,j}^y \cdot B_y^{\max} \quad (39)$$

Each of the schemes presented above maintains a reservation pool. Whenever a scheme reserves resources in a cell, the reserved resources are added to this reservation pool. In the crisis of resources, handoff services are supported by providing bandwidth from the pool. The CVP proportionate bandwidth reservation technique is more accurate, as the scheme reserves more resources in a cell for the services having higher CVP. Conversely, less amount of bandwidth (i.e., resources) is reserved in the cell if the CVP of the MU to arrive at the cell is small.

CALL ADMISSION CONTROL SCHEMES

Call admission control (CAC) schemes are used to decide whether an incoming call (new or handoff) is allowed to be admitted into a given cell for the network services or not. Admission control is necessary for the real-time continuous media services due to the following reasons. Firstly, the amount of resources requested by the service may not match with the level of resources available in the cell at the time of originating the service. So, the network checks the availability of resources while admitting a service in its CAC policy. Secondly, when a service is admitted into the network, it can be assumed that a contact is made between the service and the network in which the network guarantees to provide the required resources to the service during its span of life. So, to ensure the provision of providing the contact resources to the ongoing services, the network needs to implement a

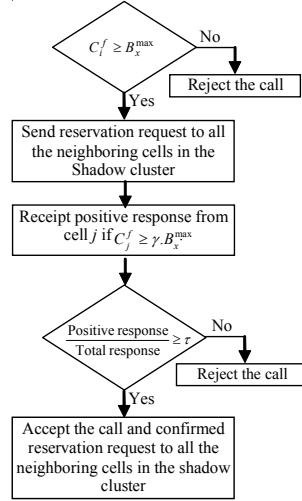
CAC policy while admitting a new or migrating (i.e., handoff) service.

The conventional CAC schemes (e.g., Gaasvik et al., 1991; Hong & Rappaport, 1986) along with many recent CAC policies (e.g., Wu et al., 2002; Kown et al., 2003; Oliviera et al., 1998) make a service admission decision solely based on the free resources (also the reserved resources for the handoff services) in the cell and the amount of desired resources requested by the service. These schemes consider only the single-class services and assume that the services are non-adaptive, having a fixed amount of resource demand during its lifetime. According to these schemes, if the free bandwidth in the cell is greater than or equal to the desired bandwidth of the newly arriving service, then the service is accepted; otherwise it is rejected. A handoff service is accepted only when the sum of the free and the reserved bandwidth can meet the demand of the bandwidth desired by the service. A similar scheme is proposed by Hou and Fang (2001) and Fang and Zhang (2002), known as the new call bounding scheme, which instead of reserving resources explicitly, uses a predefined threshold to control the admission into the cell. If the number of new services in a cell exceeds the predefined threshold, then the service is dropped; otherwise it is admitted. The handoff services are rejected only when all the channels in the cell are used up. Ramjee, Towsley, and Nagarajan (1997) proposed a CAC scheme, known as the new call thinning scheme, where admission decision for the new services are taken based on the number of busy channels—that is, the number of existing services in the cell. According to this scheme, when the number of busy channels in a cell is k , an arriving new service will be admitted with the probability of β_k ($0 \leq \beta_k \leq 1$) where $\beta_1 \geq \beta_2 \geq \dots \geq \beta_c$. An arriving handoff service is dropped when there is no channel available in the cell to accommodate it. The new call thinning scheme becomes the conven-

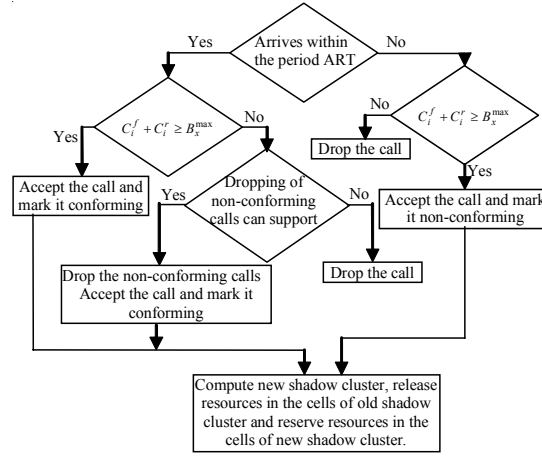
tional GC scheme with $((c - m)/c)\%$ channel reservation if $\beta_1 = \beta_2 = \dots = \beta_m = 1$ and $\beta_{m+1} = \beta_{m+2} = \dots = \beta_c = 0$. All the schemes presented above do not consider the number of MUs in the neighbouring cell having the possibility of visiting the current cell, or the MUs in the current cell having the possibility of visiting the neighbouring cells.

Most of the recent schemes (e.g., Aljadhari & Znati, 2001; Levine et al., 1997; Yu & Leung, 2002) consider the distributed approach for their CAC policies in which CAC decision is made based on the information of the MUs both in the current and the neighbouring cells. While admitting a new service, these schemes at first check whether the service will be supported by the free bandwidth in the cell or not. If the free bandwidth in the current cell is enough to be allocated the desired bandwidth for the service, a feedback policy is implemented (i.e., need to have positive feedback from the neighbouring cells more than a given predefined threshold) before admitting a new service so that the total load in the network is under control. The scheme presented by Aljadhari and Znati (2001) implements an additional condition while making a service admission decision by dividing the services in a cell into two classes: the conforming and the non-conforming services. If a handoff service arrives in a cell within its shadow cluster inside the time-period ART, then the service is termed as the conforming service. The services arriving outside its shadow cluster or the time period ART is known as the non-conforming services. The conforming services are given higher priority over the non-conforming services and accepted even if dropping the non-conforming services. The detailed CAC algorithm presented in this scheme for the new and the handoff services are shown in Figure 12, where a neighbouring cell sends a positive response if at least $\gamma\%$ of the desired bandwidth can be reserved in the cell. A new service is

Figure 12. A distributed CAC policy designed based on the conforming and the non-conforming services



CAC policy for new call



CAC policy for handoff call

accepted if at least $\tau\%$ of the neighbouring cells send the positive responses.

The aforementioned schemes, however, do not consider the adaptability of the multimedia services—that is, the services can tolerate the fluctuations in the availability of network resources by degrading their QoS within the acceptable range. The CAC policies in some schemes (e.g., El-Kadi et al., 2002; Malla, El-Kadi, Olariu, & Todorova, 2003; Kown et al., 2003; Islam & Murshed, 2004) are designed considering the tolerance of the bandwidth fluctuations supported by the multimedia services such that the resources are borrowed to admit more new and handoff services. In these schemes, each service x needs to specify its desired bandwidth $B_x^{\max}$ and minimum bandwidth $B_x^{\min}$ while requesting for admission. An expected bandwidth $B_x^{\exp}$ is estimated (see El-Kadi et al., 2002; Malla et al., 2003) by assuming a local network parameter f ($0 < f < 1$) as follows:

$$B_x^{\max} - B_x^{\min} = f(B_x^{\max} - B_x^{\exp}) \quad (40)$$

Islam and Murshed (2004) defined $B_x^{\exp}$ as shown in (41) based on the assumptions that: (1) more priority should be given to the services claiming less bandwidth to restrict the users from claiming unnecessary extra resources; and (2) as the difference between the $B_x^{\max}$ and the $B_x^{\min}$ increases, more flexibility is found for allocating resources.

$$B_x^{\exp} = B_x^{\max} - \frac{(B_x^{\max} - B_x^{\min})^2}{B_x^{\max}} \quad \text{or} \quad 2B_x^{\min} - \frac{(B_x^{\min})^2}{B_x^{\max}} \quad (41)$$

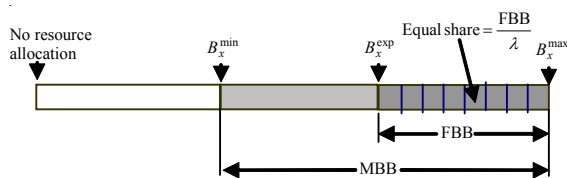
The term fair borrowable bandwidth (FBB) is given to $(B_x^{\max} - B_x^{\exp})$, while the maximum borrowable bandwidth (MBB) is the maximum amount of bandwidth which can be borrowed from x —that is, the difference between $B_x^{\max}$ and $B_x^{\min}$.

The CAC policy presented in Kown et al. (2003) admits a new or handoff service if at least $B_x^{\min}$ bandwidth can be allocated to it even if reducing the allocated bandwidth of all the existing services to its minimum value. Therefore, if the sum of the free bandwidth in the cell and the MBB of all the existing services is greater than or equal to the minimum requirements of the service (i.e., $C_i^f + \sum_k \text{MBB}_k \geq B_x^{\min}$), where k is the existing service in the cell, then the service x is accepted. This CAC scheme, however, does not provide the QoS guarantee, as the same priority level is used for both the new and the handoff services. Moreover, the long-term allocation of minimum bandwidth to any service allowed in this scheme is questionable. The schemes presented by El-Kadi et al. (2002) and Malla et al. (2003), however, solve the problem by providing more priority to the handoff services while admitting into the cell. Moreover, the schemes treat the C1 and the C2 services differently, as the C1 services act in online interactive mode while the C2 services are off-line and do not include any human interaction. The scheme proposed by El-Kadi et al. (2002) divides the FBB of each service into a predefined number of equal borrowable units, as shown in Figure 13. A new service is accepted if at least the expected bandwidth can be allocated to it, even if borrowing one unit of borrowable bandwidth from each of the existing services. A C2 service is accepted if the free bandwidth plus one borrowable bandwidth

unit from each existing service is greater than or equal to zero. On the other hand, if the minimum required bandwidth can be provided from the free and the reserved bandwidth in the cell, even if borrowing one borrowable unit of bandwidth from each of the existing services, then a C1 handoff service is accepted. The CAC scheme proposed by Malla et al. (2003) is very similar to the above scheme, except for the following two modifications. First, instead of borrowing only one borrowable bandwidth, the scheme allows you to borrow up to the FBB amount of bandwidth from each existing service while admitting a new or handoff service. Second, while admitting a new service, the system calculates average bandwidth per service in the cell by dividing the cell capacity by the number of services in the cell including the new ones. If the expected bandwidth of the service is not less than the average bandwidth per service, then the service is not allowed to be admitted in the network. This condition is due to providing more priority to the smaller services. The scheme, however, may reject services even if wasting the free bandwidth in the cell, which is sufficient enough to accommodate the service. For example, assume that there are already nine services in a cell having a capacity of 70 BUs. A new service demanding 8 BUs bandwidth will be rejected, as its demanded bandwidth is more than the average bandwidth (i.e., $70/10 = 7$ BUs) even if the cell has 10 BUs of free bandwidth.

The scheme presented by Islam and Murshed (2005) further improves the CAC strategy. The scheme suggests suspending the C2 services temporarily in the congested period to accommodate more C1 services. On the other hand, to ensure the better off-line service transmission and increase the utility of reserved resources, the idle bandwidth in the reservation pool can temporarily be borrowed by the C2 services until any C1 handoff service needs it. Due to the fully elastic nature of off-line service (as

Figure 13. Bandwidth distribution in a cell with division of FBB into equal share



the off-line service can tolerate bandwidth fluctuations from its desired rate to the minimum at zero), a C2 service is accepted if any amount of bandwidth greater than zero can be provided, even if temporarily borrowing from the reservation pool. Once a C2 service is accepted, it continues until the whole message is transferred—that is, the service never drops but may be suspended temporarily in the congested period due to the resource shortage in the cell. Any further classification of C2 services into the new or the handoff has no meaning, although the transmitting device may migrate from one cell to one of its neighbours because the C2 services are transmitted in the non-interactive off-line mode. According to the scheme, a C1 new service is accepted if the sum of the free bandwidth in the cell and the borrowing FBB from each of the existing services is greater than or equal to the expected bandwidth of the service. However, a best attempt is made to accept a C1 handoff service as it has the highest priority. A C1 handoff service is accepted even if borrowing the MBB amount of bandwidth from each of the existing services when the sum of the free and the reserved bandwidth cannot fulfil the minimum requirement. A C1 handoff service is, therefore, dropped if the sum of the free bandwidth, the reserved bandwidth, and the MBB amount of borrowed bandwidth from each of the existing services cannot accommodate it. Note that as the minimum required bandwidth for all C2 services is zero, borrowing the MBB amount of bandwidth from C2 services ultimately suspends the service.

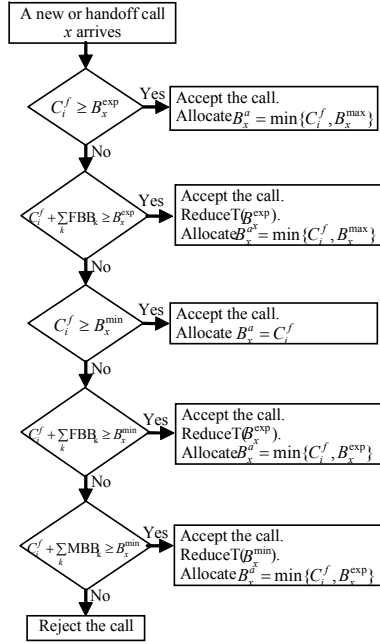
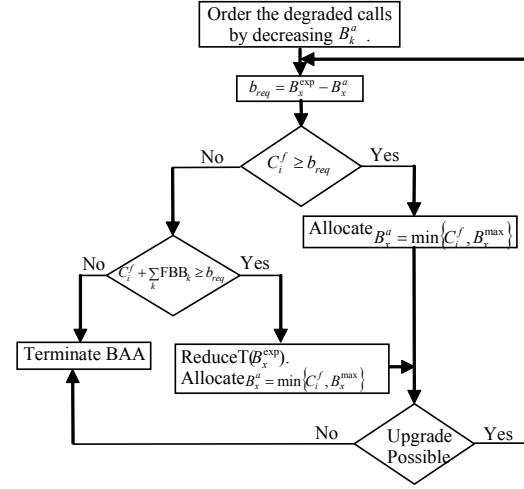
FAIR RESOURCE ALLOCATION POLICIES

Once a service (new or handoff) is accepted, the bandwidth in the cell may need to be redistributed to ensure the fair allocation of re-

sources among the competing services. Fair bandwidth allocation means the sharing of resources is achieved between the competing services based on some fairness strategy to protect the well-behaved from the aggressive applications. However, it is an extremely difficult and confusing task to allocate resources fairly among the different services in a cell, as the term “fairness” is still undefined. The proportional fairness (e.g., Mo & Walrand, 2000) and the max-min fairness (e.g., ATM Forum Technical Committee, 1996) are the two popular and widely used, but not universally accepted, fairness policies used in wired networks. Unfortunately, bandwidth constraints differ significantly in wireless networks. For an instance, in wired networks, only those users sharing the same link contend with each other in terms of bandwidth usage. However, in wireless networks, all the users in the vicinity can contend even if they are using the different wireless links. This is because transmission from one user reaches every one else in the neighbourhood.

A number of fair resource allocation schemes (e.g., El-Kadi et al., 2002; Malla et al., 2003, Islam & Murshed, 2004) have been proposed for sharing resources among the services in a cell of the mobile network. All such schemes use the techniques of borrowing bandwidth from the existing services based on some fairness principles. El-Kadi et al. (2002) proposed a rate-based borrowing (RBB) scheme by extending the proportional fairness scheme for redistributing the resources among the services in a cell in the mobile network. The max-min fairness scheme was modified by Malla et al. (2003), where the fairness criteria was used to give more priority to the smaller services than the larger ones. The idea was further enhanced in the min-max fairness scheme proposed by Islam and Murshed (2004). The different available techniques for the fair allocation of resources are presented in Figure 14.

Figure 14. Bandwidth adaptation algorithm


 Bandwidth adaptation algorithm
for the reduction process

 Bandwidth adaptation algorithm
for the expansion process

Bandwidth Adaptation Algorithms for the Fair Resource Distribution

Kown et al. (2003) proposed a bandwidth adaptation algorithm (BAA) which seeks to minimize the number of services with lower-than-the-expected bandwidth B_x^{exp} . The algorithm performs the bandwidth adaptation in two ways: the reduction process and the expansion process. The reduction process is applied to the case where a new or handoff service arrives in an overloaded cell. Depending on the situation, the BAA accommodates the incoming service by redistributing the bandwidth among the existing services—that is, reducing the bandwidth of all the existing services to their expected or minimum level, if needed. The expansion process may increase the bandwidth allocated to the remaining existing services when any one

existing service terminates or hands-off to the neighbouring cell. The detailed BAAs for both the reduction and the expansion process are presented in Figure 14. The operation $\text{ReduceT}(B_x^{\text{exp}})$ reduces the bandwidth of the existing services from the amount currently allocated to the expected level, and $\text{ReduceT}(B_x^{\text{min}})$ reduces the bandwidth to the minimum level.

The BAA scheme presented above may unnecessarily drop the bandwidth of all the existing services abruptly to either the expected or the minimum level. This draws considerable attention to all of the active users. Moreover, the scheme may waste free bandwidth in the cell while some services are getting the poor QoS.

The Rate-Based Borrowing Scheme

The rate-based borrowing (RBB) scheme (see El-Kadi et al., 2002) allocates and redistributes resources among different services in a cell according to the well-known proportional fairness scheme (see Kelly, Maulloo, & Tan, 1998; Mo & Walrand, 2000). This scheme estimates the expected bandwidth B_x^{exp} for each service x as shown in 40 and divides the FBB into a predefined number (say, λ) of equal borrowable units, as shown in Figure 13. A cell is said to be operated at level L ($0 \leq L \leq \lambda$) when all of its existing services have had L borrowable bandwidth units (or more) borrowed from them. When a new service (either C1 or C2) requests admission into the network in a cell operating at level L , the scheme first attempts to accept the service allocating the amount of bandwidth equal to its desired bandwidth minus L borrowable units (i.e., $B_x^{\text{max}} - (L/\lambda)FBB$), if available. Otherwise, the scheme tests and tries to accept the service by borrowing one equal share from each of the existing services and shifting the cell operating point from level L to $L+1$, if $L \neq \lambda$. If borrowing one unit is not sufficient to accommodate the service, then the service is rejected. Note that $L = \lambda$ implies that no borrowing is allowed—that is, it is not possible to shift the cell operating point from level L to $L+1$.

After admitting a handoff service x , the scheme allocates the bandwidth between B_x^{max} and B_x^{min} from the free and the reserved bandwidth. However, if the sum of the free and the reserved bandwidth is less than the minimum required bandwidth B_x^{min} , then one unit is borrowed from each of the existing services to allocate to the service x the maximum required bandwidth. Since for any C2 service x , $B_x^{\text{min}} = 0$, the C2 handoff services are accepted if any bandwidth greater than zero can be allocated to

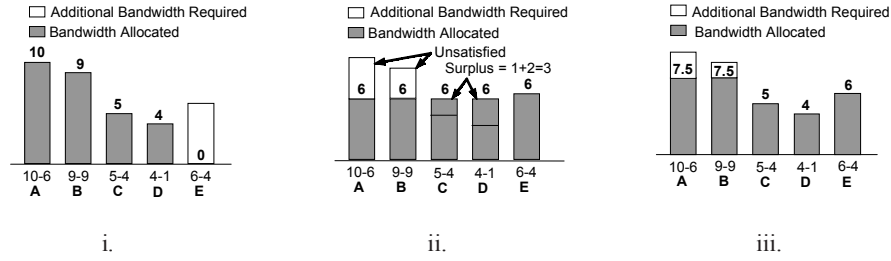
it even if borrowing one unit from each of the existing services.

The scheme does not maintain the operating point properly, due to accepting the handoff services with the allocation of the minimum bandwidth while the other existing services get sufficient resources. Moreover, the scheme unfairly allocates a large amount of bandwidth to some services (see Malla et al., 2003), as well as treats the real-time multimedia applications and non-real-time data traffic in the same way while admitting a new service.

The Max-Min Fairness Scheme

Malla et al. (2003) presented a bandwidth allocation scheme to redistribute the wireless resources among the different services in a cell using the well-known classic max-min fairness (see Mo & Walrand, 2000; Bertsekas & Gallager, 1992; Hahne, 1991) policy. An allocation is max-min fair if there is no way to give more bandwidth to a connection without decreasing the allocation of a connection of lesser or equal bandwidth. According to this scheme, if a cell has b units of bandwidth available for allocation and there are n candidate services, then each is guaranteed its equal share of b/n units. Any service requiring at most their equal share is guaranteed their desired bandwidth and referred to as a satisfied service. For instance, assume that there are n services in a cell and m satisfied services at a stage of the resource redistribution process. The total bandwidth requested by these m satisfied services is S units. The surplus bandwidth, $R = mb/n - S$, is then allocated equally among the remaining unsatisfied services, each of which is receiving $R/(n-m)$ units. After getting the equal share of surplus bandwidth, if a service receives more than or equal to its desired amount, then the service becomes satisfied and the residual bandwidth is added to the surplus bandwidth pool.

Figure 15. Distribution of bandwidth (BW) among different services according to the max-min fairness. (i) Existing distribution of bandwidth (ii) Equal share is given to every service in first step of redistribution (iii) In 2nd step; surplus bandwidth is equally redistributed among all unsatisfied services and service B will be dropped finally.



The process continues until all the services are satisfied or the surplus bandwidth becomes zero.

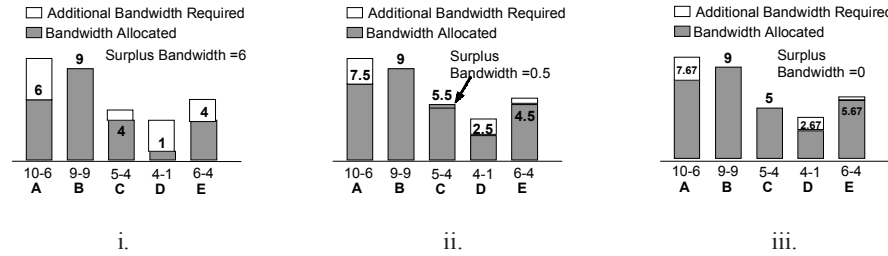
The max-min fairness scheme may terminate one or more existing services unfairly during its redistribution process. An example of presenting such an unfair termination of service is shown in Figure 15. Assume that there are four services currently in a cell: A(10-6), B(9-9), C(5-4), and D(4-1), each of which has been allocated its desired bandwidth, as shown in Figure 15(i), where the nomenclature S(R-M) means the service S has maximum and minimum bandwidth requirements of R units and M units respectively. Let the cell capacity be 30 units. If a new service E(6-4) arrives in the cell, the service will be accepted since its required bandwidth is less than or equal to the average bandwidth per service ($30/5 = 6$ units). In the start of the bandwidth redistribution process, each service will get six units, as shown in Figure 15(ii), and the services A and B become unsatisfied as they do not receive their desired bandwidth. Conversely, 3 units surplus bandwidth is found from the service C and D (1 unit from C and 2 units from D) as they have more than their desired amount. So, in the first step of the surplus bandwidth redistribution process, the surplus bandwidth is distributed equally among the two unsatisfied services (A and B).

Finally, the services A, B, C, D, and E will receive 7.5, 7.5, 5, 4, and 6 units respectively, as shown in Figure 15(iii). Unfortunately, as the service B has the minimum bandwidth requirement of 9 units, it will be terminated unfairly.

The Min-Max Fairness Scheme

The min-max fairness scheme (see Islam & Murshed, 2004) is proposed with the aims of ensuring at least the minimum bandwidth allocation to all the services and not dropping a service if the minimum resources can be provided. Besides considering the above principles, the scheme distributes bandwidth as evenly as possible among the services without unduly reducing the throughput. The basic premise of the min-max fairness scheme is to allocate the desired bandwidth to the new or handoff services if the free bandwidth (plus the reserve bandwidth for handoff services) is sufficient enough to do so. Otherwise, the scheme firstly allocates the minimum required bandwidth to all the services in an overloaded cell, as it is unfair to allocate more bandwidth to the other services, while a service has no bandwidth to survive. The surplus bandwidth is then equally distributed between all the unsatisfied services. If a service cannot utilize its allocated bandwidth because of constraints elsewhere, the

Figure 16. Distribution of bandwidth (BW) among different services according to the min-max fairness. (i) Minimum bandwidth is allocated to each service; (ii) Equal share is given to the unsatisfied services in first step of redistribution (iii) In 2nd step; surplus bandwidth is equally redistributed among all unsatisfied services.



residual bandwidth is added to the surplus bandwidth. The process of distributing the surplus bandwidth among the unsatisfied services is repeated until all the services in the cell are satisfied or there is no surplus bandwidth available.

Figure 16 shows an example of the bandwidth redistribution process used in the min-max fairness scheme for the same scenario shown in Figure 15(i). The min-max fairness scheme first allocates the minimum bandwidth to all the services in the cell and has 6 units of surplus bandwidth. In first iteration, the bandwidth is equally distributed among the four unsatisfied services (A, C, D, and E), as shown in Figure 16(ii). The residual bandwidth (0.5 units) from the service C is then collected to the surplus bandwidth. In the second iteration, the surplus bandwidth is equally distributed among the unsatisfied services A, D, and E. The process terminates after the second redistribution of surplus bandwidth, as the surplus bandwidth at this stage becomes zero.

CONCLUSION AND FUTURE DIRECTIONS

The accommodation of adaptive multimedia services with the diverse quality of service (QoS) requirements required for mobile busi-

ness poses many new challenges in the design of resource management strategies, such as how to determine the efficient amount of resources needing to be reserved for migrating (i.e., handoff) services, what should be the time window for reserving resources, how to determine whether a given neighbouring cell needs to reserve resources or not, how to efficiently allocate the available resources to optimize channel utilization and ensure fairness, as well as how to ensure whether a service (new or handoff) seeking admission into the cell is to be accepted or not. This chapter outlines all of these resource management strategies so that an efficient network can be developed for facilitating the communications required for the mobile business.

The users' mobility is one of the main influential and differential factors for designing the resource management strategies for a mobile network. Although the conventional strategies do not consider the mobility information while managing their scarce resources, those were developed to support the voice services only. The use of mobility information is very essential for designing the call admission control, the resource reservation, and the resource allocation strategies for the adaptive multimedia services which can tolerate the transient fluctuations in their QoS assurance. Many schemes

have been designed for predicting the users' future mobility from the historical databases and use this predicted mobility information for making the decision to admit a new or handoff service, reserving and allocating resources in each cell. These schemes however are very complex, time consuming, and suffer from the lack of accuracy due to considering the huge databases and ignoring the mobile users' present status. Some other recent schemes use the combination of key mobility parameters such as distance, direction, and speed, as well as some statistical measurements based on the past observations, and represent the nature of road networks and geographical environments. These schemes, however, present feasible resource management strategies which can be used for efficient resource management in future B3G/4G mobile networks.

Although the use of the mobility parameters with the statistical measures and the call duration information for estimating the neighbouring cell visiting probability is a good technique, analysis of traffic trends can also help a lot for estimating the future position of any mobile user and reserving resources for the migrating services. The changes in traffic flow normally occur in a cyclic way and highly influence the prediction process of future mobility. On the other hand, it is quite difficult to ensure the fair distribution of resources among the competing services. The utility of resources is different to the different users based on their service categories. Therefore, for fair distribution of resources, the services should be categorised according to their utility and bandwidth requirements so that the resources can be provided to the services having better utility of it.

REFERENCES

- Acampora, A. S., & Naghshineh, M. (1994). An architecture and methodology for mobile-executed handoff in cellular ATM networks. *Journal on Selected Areas in Communications*, 12, 1365-1375.
- Aljadhari, A., & Znati, T. F. (2001). Predictive mobility support for QoS provisioning in mobile wireless environments. *IEEE Journal on Selected Areas in Communications*, 19(10), 1915-1930.
- ATM Forum Technical Committee. (1996). *Traffic management specification* (v. 4.0).
- Bertsekas, D., & Gallager, R. (1992). *Data networks*. Englewood Cliffs, NJ: Prentice-Hall.
- Choi, S., & Shin, K. G. (2002). Adaptive bandwidth reservation and admission control in QoS-sensitive cellular networks. *IEEE Transactions on Parallel and Distributed Systems*, 13(9), 882-897.
- Dreyfus, S. E. (1969). An appraisal of some shortest-path algorithms. *Operational Research*, 17, 395-412.
- El-Kadi, M., Olariu, S., & Abdel-Wahab, H. (2002). A rate-based borrowing scheme for QoS provisioning in multimedia wireless networks. *IEEE Transactions on Parallel and Distributed Systems*, 13(2), 156-166.
- Fang, F., Chlamtac, I., & Lin, Y.-B. (1998). Channel occupancy times and handoff rate for mobile computing and PCS networks. *IEEE Transactions on Computers*, 47, 679-692.
- Fang, Y., & Zhang, Y. (2002). Call admission control schemes and performances analysis in wireless networks. *IEEE Transactions on Vehicular Technology*, 51(2), 371-382.
- Gaasvik, P., Corneffjord, M., & Svensson, V. (1991). Different methods of giving priority to handoff traffic in a mobile telephone system with directed retry. *Proceedings of the 41st IEEE Vehicular Technology Conference*:

Gateway to the Future Technology in Motion (pp. 549-553).

Hahne, E. (1991). Round-robin scheduling for max-min fairness in data networks. *IEEE Journal on Selected Areas in Communications*, 9, 1024-1039.

Hong, D., & Rappaport, S. (1986). Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and non-prioritized handoff procedures. *IEEE Transactions on Vehicular Technology*, 35(3), 77-92.

Hou, J., & Fang, Y. (2001). Mobility-based call admission control schemes for wireless mobile networks. *Wireless Communications and Mobile Computing*, 1(3), 269-282.

Islam, M. M., & Murshed, M. (2004). Novel velocity and call duration support for QoS provision in mobile wireless networks. *IEEE Wireless Communications*, 11(5), 22-30.

Islam, M. M., Murshed, M., & Dooley, L. S. (2003). Influence of velocity on resource reserving directional probability function estimation for mobile multimedia communications. *Proceedings of the International Symposium on Intelligent Signal Processing and Communications (ISPACS)*, Awaji Island, Japan (pp. 64-69).

Islam, M. M., Murshed, M., & Dooley, L. S. (2004). A novel velocity-dependant directional probability function based call admission control schemes in wireless multimedia communications. *Proceedings of the 23rd IEEE International Performance, Computing, and Communications Conference (IPCCC)*, Phoenix, AZ (pp. 77-84).

Islam, M. M., Murshed, M., & Dooley, L. S. (2004). A novel mobility support resource reservation and call admission control scheme for quality-of-service provision in wireless multi-

media communications. *Proceedings of the IEEE International Conference on Communications (ICC)*, Paris, France (Vol. 1, pp. 48-52).

Islam, M. M., Murshed, M., & Dooley, L. S. (2003). New mobility based call admission control with on-demand borrowing scheme for QoS provisioning. *Proceedings of the International Conference on Information Technology (ITCC)*, Las Vegas, NV (pp. 263-267).

Islam, M. M., Murshed, M., & Dooley, L. S. (2004). Enhanced cell visiting probability for QoS provisioning in mobile multimedia communications. *Proceedings of the International Conference on Information Technology (ITCC)*, Las Vegas, NV (Vol. 2, pp. 258-262).

Islam, M. M., & Murshed, M. (2005). An adaptive min-max scheme for resource allocation in cellular multimedia networks. *European Transactions on Telecommunications* (accepted for publication).

Kampen, N. G. V. (1981). *Stochastic processes in physics and chemistry*. Amsterdam: North Holland.

Kanakia, H., Mishra, P. P., & Reibman, A. (1996). An adaptive congestion control scheme for real-time packet video transport. *IEEE/ACM Transactions on Networking*, 3.

Kelly, F., Maulloo, A., & Tan, D. (1998). Rate control for communication networks: Shadow price proportional fairness and stability. *Journal of the Operational Research Society*, 49, 237-252.

Kown, T., Choi, Y., Bisdikian, C., & Naghshineh, M. (2003). QoS provisioning in wireless/mobile multimedia networks using an adaptive framework. *Wireless Networks*, 9, 51-59.

Levine, D. A., Akyildiz, I. F., & Naghshineh, M. (1997). A resource estimation and call admission algorithm for wireless multimedia

- networks using the shadow cluster concepts. *IEEE/ACM Transactions on Networking*, 5(1), 1-12.
- Li, B., Lin, C., & Chanson, S. (1998). Analysis of a hybrid cutoff priority scheme for multiple classes of traffic in multimedia wireless networks. *ACM/Baltzer Journal on Wireless Networks*, 4(4), 279-290.
- Liu, G. Y., & Maguire, G. Q. (1995). A predictive mobility management algorithm for wireless mobile computation and communication. *Proceedings of the IEEE International Conference on Universal Personal Communications* (pp. 268-272).
- Liu, T., Bahl, P., & Chlamtac, I. (1998). Mobility modeling, location tracking, and trajectory prediction in wireless ATM networks. *IEEE Journal on Selected Areas in Communications*, 16(6), 922-936.
- Malla, A., El-Kadi, M., Olariu, S., & Todorova, P. (2003). A fair resource allocation protocol for multimedia wireless networks. *Proceedings of the IEEE Transactions on Parallel and Distributed Systems*, 14(1), 63-71.
- McMillan, D. (1995). Delay analysis of a cellular mobile priority queuing system. *IEEE/ACM Transactions on Networking*, 3, 310-319.
- Mo, J., & Walrand, J. (2000). Fair end-to-end window-based congestion control. *IEEE/ACM Transactions on Networking*, 8(5), 556-567.
- Naghshineh, M., & Schwartz, M. (1996). Distributed call admission control in mobile/wireless networks. *IEEE Journal on Selected Areas in Communications*, 14.
- Oliviera, C., Kim, J., & Suda, T. (1998). An adaptive bandwidth reservation scheme for high speed multimedia wireless networks. *IEEE Journal on Selected Areas in Communications*, 12, 858-874.
- Posner, E.C., & Guerin, R. (1985). Traffic policies in cellular radio that minimize blocking of handoff calls. *Proceedings of the 11th ITC*, Kyoto, Japan (pp. 294-298).
- Ramjee, R., Nagarajan, R., & Towsley, D. (1996). On optimal call admission control in cellular networks. *Proceedings of IEEE INFOCOM*, San Francisco (Vol. 1, pp. 43-50).
- Ramjee, R., Towsley, D., & Nagarajan, R. (1997). On optimal call admission control in cellular networks. *Wireless Networks*, 3, 29-41.
- Tabbane, S. (1995). An alternative strategy for location tracking. *IEEE Journal on Selected Areas on Communications*, 13.
- Tekinay, S., & Jabbari, B. (1992). A measurement-based prioritization scheme for handovers in mobile cellular networks. *IEEE Journal on Selected Areas in Communications*, 10(8), 1343-1350.
- Van, V.P., & Glisic, S. (2001). Performance analysis of queuing schemes for priority handoff and call admission control in mobile cellular radio networks. *Proceedings of the International Conference on Communications (ICC)* (Vol. 7, pp. 2291-2295).
- Van, V.P., & Glisic, S. (2001). An analytical modeling for a class of queuing priority handoff and call admission control schemes in micro-cellular PCNs. *Proceedings of the IEEE Vehicular Technology Conference (VTC)* (Vol. 2, pp. 901-905).
- Vickers, B.J., Lee, M., & Suda, T. (1997). Feedback control mechanisms for real-time multipoint video services. *IEEE Journal on Selected Areas in Communications*, 15.
- Wu, S., Wong, K.Y.M., & Li, B. (2002). A dynamic call admission policy with precision QoS guarantee using stochastic control for

mobile wireless networks. *IEEE Transactions on Networking*, 10(2), 257-271.

Yu, F., & Leung, V. (2002). Mobility-based predictive call admission control and bandwidth reservation in wireless cellular networks. *Computer Networks*, 38, 577-589.

Ziv, J., & Lempel, A. (1978). Compression of individual sequences via variable rate coding. *IEEE Transactions on Information Theory*, 24, 530-536.

Zhou, X., Chen, R., & Gao, C. (2001). An elliptical shadow algorithm for motion prediction and resource reservation in wireless cellular networks. *Proceedings of the 10th International Conference Computer Communications and Networks (ICCN)* (pp. 234-239).

Section V

Security

Chapter XXV

Secure Agent Roaming for Mobile Business

Sheng-Wei Guan
Brunel University, UK

ABSTRACT

M-commerce, a new way to conduct business, is gaining more and more popularity due to the wide use of the Internet. Despite its rapid growth, there are limitations that hinder the expansion of m-commerce. The primary concern for online shopping is security. Due to the open nature of the Internet, personal financial details necessary for online shopping can be stolen if sufficient security mechanism is not put in place. How to provide the necessary assurance of security to consumers remains a question mark despite various past efforts. Another concern is the lack of intelligence in locating the correct piece of information. The Internet is an ocean of information depository. It is rich in content but lacks the necessary intelligent tools to help one locate the correct piece of information. Intelligent agent, a piece of software that can act intelligently on behalf of its owner, is designed to fill this gap. However, no matter how intelligent an agent is, its functionality is limited if it remains on its owner's machine and does not have any roaming capability. With the roaming capability, more security concerns arise. In response to these concerns, SAFE, Secure roaming Agent For E-commerce, is designed to provide secure roaming capability to intelligent agents.

INTRODUCTION

The introduction of the Internet is probably one of the most significant revolutions of the 20th century. With a simple click, one can connect to almost every corner of the world thousands of kilometers away. This presents a great oppor-

tunity for mobile commerce (m-commerce). Despite its many advantages over traditional commerce, m-commerce has not taken off successfully. One of the main hindrances is security.

When it comes to online transactions, security becomes the primary concern. The Internet

was developed without too much security in mind. Information flows from hubs to hubs before it reaches the destination. By simply tapping into wires or hubs, one can easily monitor all traffic transmitted. For example, when Alice uses her VISA credit card to purchase an album from Virtual CD Mall, the information about her card may be stolen if it is not carefully protected. This information may be used maliciously to make other online transactions, thus causing damage to both the card holder and the credit card company.

Besides concerns on security, current m-commerce lacks intelligence. The Internet is like the world's most complete library collections unsorted by any means. To make things worse, there is no competent librarian that can help readers locate the book wanted. Existing popular search engines are attempts to provide librarian assistance. However, as the collection of information is huge, none of the librarians are competent enough at the moment.

The intelligent agent is one solution to providing intelligence in m-commerce. But having an agent that is intelligent is insufficient. There are certain tasks that are unrealistic for agents to perform locally, especially those that require a large amount of information. Therefore, it is important to equip intelligent agents with roaming capability.

Unfortunately, with the introduction of roaming capability, more security issues arise. As the agent needs to move among external hosts to perform its tasks, the agent itself becomes a target of attack. The data collected by agents may be modified, the credit carried by agents may be stolen, and the mission statement on the agent may be changed. As a result, transport security is an immediate concern to agent roaming. SAFE transport protocol is designed to provide a secure roaming mechanism for intelligent agents. Here, both general and roaming-related security concerns are addressed carefully. Furthermore, several protocols are de-

signed to address different requirements. An m-commerce application can choose the protocol that is most suitable based on its need.

Background on Agents

There has been a lot of research in the area of intelligent agents. Some literature only proposes certain features of intelligent agents, some attempts to define a complete agent architecture. Unfortunately, there is no standardization in the various proposals, resulting in vastly different agent systems. Efforts are made to standardize some aspect of agent systems so that different systems can inter-operate with each other. In the area of knowledge representation and exchange, one of the most widely accepted standards is KQML (Knowledge Query and Manipulation Language) (Finin & Weber, 1993), developed as part of the *Knowledge Sharing Effort*. KQML is designed as a high-level language for runtime exchange of information between heterogeneous systems. Unfortunately, KQML is designed with little security considerations because no security mechanism is built to address common security concerns, not to mention specific security concerns introduced by mobile agents. Agent systems using KQML will have to implement security mechanisms on top of KQML to protect them. In an attempt to equip KQML with 'built-in' security mechanisms, Secret Agent is proposed by Thirunavukkarasu, Finin, and Mayfield (1995).

Secret Agent defines a security layer on top of KQML. Applications will have to implement some special message format in order to make use of Secret Agent. Secret Agent has a number of shortcomings and is handicapped by the design of KQML. Firstly, one requirement of Secret Agent is that every agent implementing the security algorithm must possess a key (master key). This master key is either a symmetric key or based on PKI. If the key is based on a

symmetric key algorithm, it requires each agent to have a separate key with every other agent it corresponds with. If the agent intends to communicate with another agent that it has no common pre-established master key, a central authentication server is required to generate such a key. The problems introduced are key database management, authentication server protection, and key transport/exchange security.

If the master key is based on PKI, the agent identity must be tightly tied with the key pair. This was insufficiently addressed in the design of Secret Agent, subjecting the algorithm to man-in-the-middle attack. In the SAFE transport protocol, agent identity and key pair are tightly integrated using digital certification.

Another prominent transportable agent system is Agent TCL, developed at Dartmouth College (Gray, 1997; Kotz, Marzullo, & Lauvset, 1997). Agent TCL addresses most areas of agent transport by providing a complete suite of solutions. It is probably one of the most complete agent systems under research. Its security mechanism aims at protecting resources and the agent itself. Since some existing agent systems are already very strong in this area, Agent TCL “seeks to confirm their sufficiency and either copy or redesign as appropriate” (Gray, 1997). In terms of agent protection, the author acknowledges that:

...it is clear that it is impossible to protect an agent from the machine on which the agent is executing...it is equally clear that it is impossible to protect an agent from a resource that willfully provides false information. (Gray, 1997)

As a result, the author:

...seeks to implement a verification mechanism so that each machine can check

whether an agent was modified unexpectedly after it left the home machine. (Gray, 1997)

In other words, it addresses agent integrity and provides a certain level of traceability to the agents. The other areas of security—like non-repudiation, verification, and identification—are not carefully addressed.

Compared with the various agent systems discussed above, SAFE is designed to address the special needs of m-commerce. The other mobile agent systems are either too general or too specific to a particular application. By designing SAFE with m-commerce application concerns in mind, the architecture will be suitable for m-commerce applications. The most important concern is security, as already discussed. Due to the nature of m-commerce, security becomes a prerequisite for any successful m-commerce application. Other concerns are mobility, efficiency, and interoperability. In addition, the design allows certain flexibility to cater to different application needs.

GENERAL AGENT TRANSPORT

As a prerequisite, each SAFE entity must carry a digital certificate issued by SAFE Certificate Authority, or SCA. In this way, each agent, agent owner, and host will carry its own unique digital certificate. The certificate itself is used to establish the identity of a SAFE entity. Because the private key to the certificate has signing capability, this allows the certificate owner to authenticate itself to the SAFE community. An assumption is made that the agent private key can be protected by function hiding (Thomas, 1998; other techniques are also discussed in Bem (2000) and Westhoff (2000), but will not be elaborated in this chapter).

From the host’s viewpoint, an agent is a piece of foreign code that executes locally. In

order to prevent a malicious agent from abusing the host resources, the host should monitor the agent's usage of resources (e.g., computing resources, network resources). Agent receptionist will act as the middleman to facilitate and monitor agent communication with external party.

General Message Format

In SAFE, agent transport is achieved via a series of message exchanges. The format of a general message is as follows:

SAFE Message = Message Content + Timestamp + Sequence Number + MD(Message Content + Timestamp + Sequence Number) + Signature(MD)

The main body of a SAFE message comprises message content, a timestamp, and a sequence number. The message content is defined by individual messages.

A timestamp contains the issue and expiry time of the message. If the message arrives before the issue time of the message or after the expiry time of the message (assuming there is no time lag between the sender and receiver), the recipient should generate an alert to the message sender as well as the recipient's administrator. The default message lifetime (time duration between issue time and expiry time) is set in the SAFE community. However, individual entities can choose to set a different message lifetime based on their needs. The local setting will overwrite the general setting by SAFE. The general guideline is that the duration must be longer than the maximum tolerable time for message exchange to complete, but less than maximum tolerable agent transport time.

To further prevent replay attack, message exchanges between entities during agent trans-

port is labeled according to each transport session. A running sequence number is included in the message body whenever a new message is exchanged. In this way, if a message is lost during transmission or an additional message is received, the recipient will be able to detect it.

In order to protect the integrity of the main message body, a message digest is appended to the main message. The formula of the message digest is as follows:

Message Digest = MD5(SHA(message_body) + message_body)

The message digest alone is not sufficient to protect the integrity of a SAFE message. A malicious hacker can modify the message body and recalculate the value of message digest using the same formula and produce a seemingly valid message digest. To ensure the authenticity of the message, a digital signature on the message digest is generated for each SAFE message. In addition to ensuring message integrity, the signature serves as a proof for non-repudiation as well.

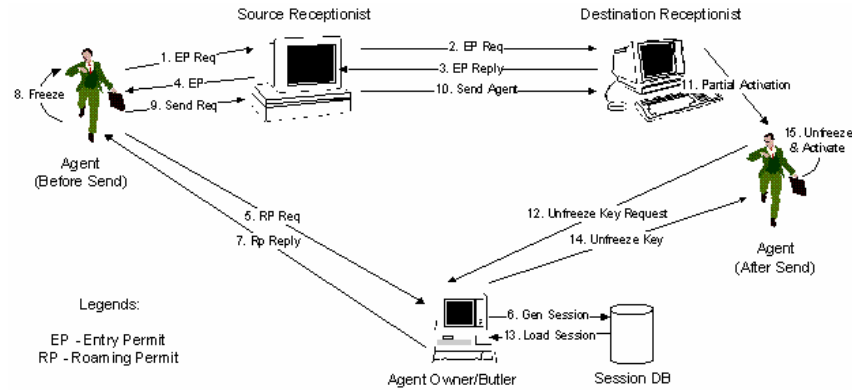
If the message content is sensitive, it can be encrypted using a symmetric key algorithm (e.g., Triple DES). SAFE does not provide a general key exchange protocol for general messages. The secret key used for encryption will have to be decided at a higher level.

To cater for different application concerns, three transport protocols are proposed: supervised agent transport, unsupervised agent transport, and bootstrap agent transport. These three protocols will be discussed in the following sections in detail.

Supervised Agent Transport

Supervised agent transport is designed for applications that require close supervision of

Figure 1. Supervised agent transport



agents. Under this protocol, an agent has to request a roaming permit from its owner or butler before roaming. The owner has the option to deny the roaming request and prevent its agent from roaming to undesirable hosts. Without the agent owner playing an active role in the transport protocol, it is difficult to have tight control over agent roaming.

The procedure for supervised agent transport is shown in Figure 1.

Agent Receptionist

Agent receptionists are processes running at every host to facilitate agent transport. If an agent wishes to roam to a host, it should communicate with the agent receptionist at the destination host to complete the transport protocol. Every host will keep a pool of agent receptionists to service incoming agents. Whenever an agent roaming request arrives, an idle agent receptionist from the pool will be activated to entertain the request. In this way, a number of agents can be serviced concurrently. The number of agent receptionists in the pool should be set to the maximum number of acceptable concurrent visiting agents in the host. If the number of roaming requests exceeds the number of agent receptionists, the

request will not be granted until some existing visiting agent leaves the host.

Request through Source Receptionist for Entry Permit

To initiate supervised agent transport, an agent needs to request an entry permit from a destination receptionist. Communication between visiting agent and foreign parties (other agents outside the host, agent owner, etc.) is done using an agent receptionist as a proxy. The request for entry permit is first sent to the source receptionist. The request contains the requesting agent's digital certificate and the destination's address. The source receptionist will forward the agent's digital certificate to the destination receptionist as specified in the agent's request.

The destination receptionist can inspect the requesting agent's information by reading its digital certificate and decide whether to issue entry permit based on its own authorization policy. If the request is granted, an entry permit is generated and returned to the requesting agent. The entry permit will contain a random challenge, a serial number, a validity period, the digital certificate of the requesting agent, and a digital signature by the destination receptionist on the entry permit.

The random challenge is used to authenticate the incoming agent. Its usage will be discussed later in the discussion of supervised agent transport protocol. For bookkeeping purposes, a serial number is included in the entry permit issued by a receptionist. This number should be unique to all entry permits issued by the same receptionist. A timestamp is also part of the entry permit. Different from timestamps on general messages, the timestamp on an entry permit specifies the validity of the entry permit. It is up to each receptionist to decide how long the issued entry permit remains valid. In order to prove the authenticity of the entry permit, the issuing receptionist needs to digitally sign the entry permit.

Request for Roaming Permit

Once the source receptionist receives the entry permit from the destination receptionist, it simply forwards it to the requesting agent. The next step is for the agent to receive a roaming permit from its owner/butler. The agent sends the entry permit and address of its owner/butler to the source receptionist. Without processing, the source receptionist forwards the entry permit to the address as specified in the agent request.

The agent owner/butler can decide whether the roaming permit should be issued based on its own criteria. If the agent owner/butler decides to issue the roaming permit, it will have to generate a session number, a random challenge, and a freeze/unfreeze key pair. The roaming permit should contain the session number, random challenge, freeze key, timestamp, entry permit, and a signature on all the above from the agent owner/butler.

In order to verify that the agent has indeed reached the intended destination, a random challenge is generated into the roaming permit. A digital signature on this random challenge is required for the destination to prove its authen-

ticity. This will be discussed in greater detail later.

For the issuing of every roaming permit, a key pair is generated. A public key is included in the roaming permit for agents to encrypt or freeze its sensitive code/data during roaming. When the agent reaches the destination, it can obtain the private key (unfreeze key) from its owner to activate itself.

Like an entry permit, a roaming permit also contains a timestamp that specifies the validity of the permit. As a general guideline, the validity should be the same as that in the entry permit unless the validity specified in the entry permit is deemed inappropriate.

Since a roaming permit is issued based on the entry permit presented, the entry permit will be part of the roaming permit. In this way, a roaming permit issued to entry permit A cannot be used as a valid roaming permit to enable an agent roaming using entry permit B.

Finally, to provide non-repudiation, the agent owner/butler will digitally sign the roaming permit.

Agent Freeze

With the roaming permit and entry permit, the agent is now able to request for roaming from the source receptionist. In order to protect the agent during its roaming, sensitive function and codes inside the agent 'body' will be frozen. This is achieved using the freeze key in the roaming permit. Even if the agent is intercepted during its transmission, the agent's capability is restricted. Not much harm can be done to the agent owner/butler. To ensure a smooth roaming operation, the agent's 'life support systems' cannot be frozen. Functions that are critical to the agent's roaming capability, such as a basic communication module and an unfreeze operation module (which requires an unfreeze key to execute), must remain functional when the agent is roaming. All other functions and data not critical to agent roaming can be frozen and

subsequently activated when the agent reaches its destination.

Agent Transport

Once frozen, the agent is ready for transmission over the Internet. To activate roaming, the agent sends a request containing the roaming permit to the source receptionist. The source receptionist can optionally verify the validity and authenticity of the roaming permit. Since the roaming permit (as well as the entry permit inside it) will be inspected one more time when it reaches the destination receptionist, the inspection by the source receptionist is optional.

If the agent's roaming permit is valid, the source receptionist will transmit the frozen agent to the destination receptionist as specified in the entry permit. Once the transmission is completed, the source receptionist will terminate the execution of the original agent and make it available to other incoming agents. The involvement of the source receptionist in the transport ends here.

Agent Pre-Activation

When the frozen agent reaches the destination receptionist, it will inspect the agent's roaming permit and the entry permit (contained in the roaming permit) carefully. By doing so, the destination receptionist can establish the following:

1. The agent has been granted permission to enter the destination.
2. The entry permit carried by the agent has not expired.
3. The agent has obtained sufficient authorization from its owner/butler for roaming.
4. The roaming permit carried by the agent has not expired.

If the destination receptionist is satisfied with the agent's credentials, it will activate the agent partially and allow it to continue agent transport process.

Request for Unfreeze Key and Agent Activation

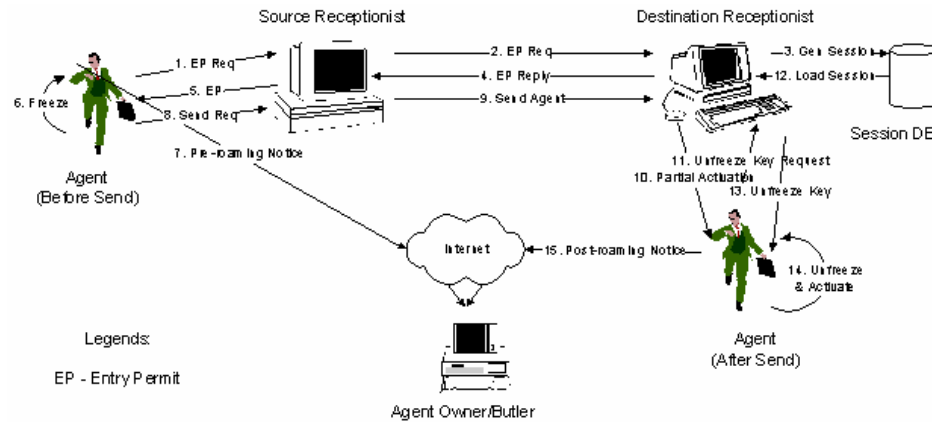
Although the agent has been activated, it is still unable to perform any operation since all sensitive codes/data are frozen. To unfreeze the agent, it has to request the unfreeze key from its owner/butler. To prove the authenticity of the destination, the destination receptionist is required to sign the random challenge in the roaming permit. The request for unfreeze key contains the session number, the certificate of destination, and the signature on the random challenge.

The agent owner/butler can verify that the agent has indeed reached the right destination by validating the signature. If the signature is valid, the agent owner/butler will retrieve the unfreeze key based on session number, encrypt it using the destination's public key, and return it to the agent.

The destination receptionist can decrypt the unfreeze key using its private key and pass the unfreeze key to the agent. Using the unfreeze key, the agent unfreezes itself. To prove to the destination host that the incoming agent is indeed the agent requesting the entry permit, the agent will use its private key to sign the random challenge in the entry permit and return it to the destination receptionist. Once this signature has been verified, the destination receptionist fully activates the agent so that it can continue its execution in the new host.

The direct agent transport process is completed.

Figure 2. Unsupervised agent transport



Unsupervised Agent Transport

Supervised agent protocol is not a perfect solution to agent transport. Although it provides tight supervision to an agent owner/butler, it has its limitations. Since the agent owner/butler is actively involved in the transport, the protocol inevitably incurs additional overhead and network traffic. This results in lower efficiency of the protocol. This is especially significant when the agent owner/butler is located behind a network with lower bandwidth, or the agent owner/butler is supervising a large number of agents. In order to provide flexibility between security and efficiency, unsupervised agent transport is proposed. The steps involved in unsupervised agent transport are shown in Figure 2.

Request for Entry Permit

In supervised agent transport, session ID and key pair are generated by the agent butler. However, for unsupervised agent transport, these are generated by the destination receptionists because agent butler is no longer online to the agents.

Pre-Roaming Notification

Unlike supervised agent transport, the agent does not need to seek explicit approval to roam from its owner/butler. Instead, a pre-roaming notification is sent to the agent owner/butler through indirect means. It serves to inform the agent owner/butler that the agent has started its roaming. The agent does not need to wait for the owner/butler's reply before roaming.

Agent Freeze

Agent freeze is very close to the same step under supervised agent transport, only that the encryption key is generated by destination instead of agent butler.

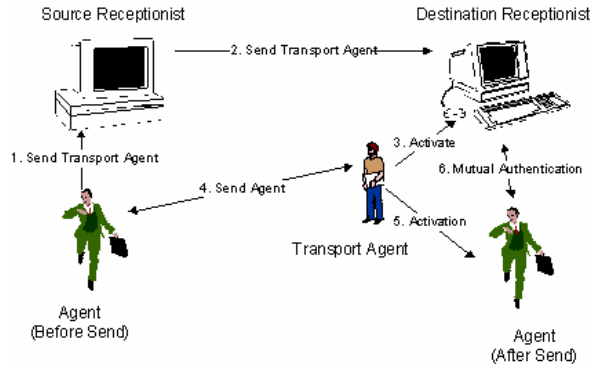
Agent Transport

This step is the same as that in supervised agent transport protocol.

Request for Unfreeze Key

The identification and verification processes are the same as in supervised agent transport,

Figure 3. Bootstrap agent transport



the exception being that the unfreeze key comes from the destination receptionist.

Agent Activation

This step is the same as that in supervised agent transport.

Post-Roaming Notification

Upon full activation, the agent must send a post-roaming notification to its owner/butler. This will inform the agent owner/butler that the agent roaming has been completed successfully. Again, this notification will take place through an indirect channel so that the agent does not need to wait for any reply before continuing with its normal execution.

BOOTSTRAP AGENT TRANSPORT

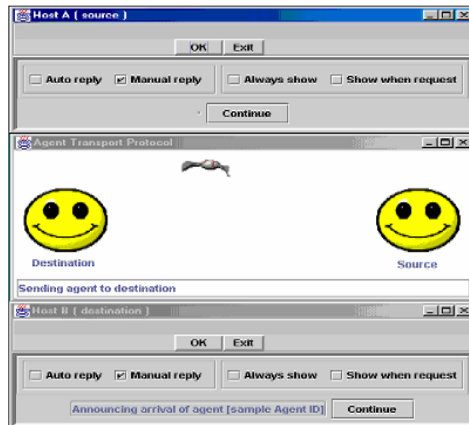
Both supervised and unsupervised agent transport make use of a fixed protocol for agent transport. The procedures for agent transport in these two protocols have been clearly defined without much room for variations. It is realized that there exist applications that re-

quire a special transport mechanism for their agents. For example, applications that involve highly sensitive content may wish to use a proprietary protocol for their agent transport. In order to allow this flexibility, SAFER provides a third transport protocol, bootstrap agent transport. Under bootstrap agent transport, agent transport is completed in two phases. The first phase is to send a transport agent to the destination using either supervised or unsupervised agent transport. In the second phase, the transport agent takes over the role of destination receptionist and continues the transport of its parent agent with its own agent transport protocols. In this way, different applications can implement their transport agents using the preferred transport mechanisms and still be able to make use of the SAFER agent transport. Bootstrap agent transport is illustrated in Figure 3.

In the first phase, the transport agent is sent to the destination receptionist using either supervised or unsupervised agent transport with some modifications. The original supervised and unsupervised agent transport requires agent authentication and destination authentication to make sure that the right agent reaches the right destination. Under bootstrap agent transport, the transmission of transport agent does not require both agent authentication and destination authentication.

Once the transport agent reaches the destination, it starts execution in a restricted environment. It is not given the full privilege as a normal agent because it has yet to authenticate itself to the destination. Under the restricted environment, the transport agent is not allowed to interact with local host services. It is only allowed to communicate with its parent until the parent reaches its destination. A maximum timeframe is imposed on the transport agent during the transmission of its parent to complete. This is to prevent the transport agent

Figure 4. Screenshot of agent transport protocol



from hacking attempts to the local host. SAFER allows individual transport agents to be customized to use any secure protocol for parent agent transmission. Concerns such as anonymity, secrecy, integrity, and so forth should be taken care of by the transport agent. If the algorithm used by the transport agent is not secure, the whole agent may be compromised. In SAFER, parent agent assumes the responsibility of making sure its transport agent uses a secure transport protocol.

When the parent agent reaches the destination, it can continue the handshake with the destination receptionist and perform mutual authentication directly. The authentication scheme is similar to that in supervised/unsupervised agent transport.

IMPLEMENTATION

To prototype the design of agent transport, the three protocols discussed above have been implemented.

The prototype is built on Windows 95/NT platform using Java (see screenshot in Figure 4). Since Java is a platform-independent lan-

guage, the prototype can be deployed to any other platform that supports JVM (Java Virtual Machine). There are a few reasons why Java is chosen as the implementation language. Firstly, the most powerful feature of Java—platform independence—makes it the ideal language for building Internet-based applications. In order to provide interoperability across multi-vendor platforms, a truly platform-independent language is desired. With Java, the prototype can be built once and run anywhere on other platforms.

Furthermore, the garbage collector feature of Java significantly reduces the programming effort and allows developers to concentrate on programming logic rather than taking care of memory. Unlike some other languages such as C/C++, Java VM manages memory automatically through its garbage collector.

Another feature that benefits the prototyping is thread-safe. Java language makes it easy to develop multi-thread applications. Threading is taken care of by JVM so that applications using Java threading is automatically thread-safe. In other languages, extra effort is needed to ensure the program runs normally under multi-thread scenario.

As a first step in the prototyping, unsupervised agent transport has been implemented. Two agent receptionists are set up in different hosts simulating the source host and destination host. An agent carrying certain functions is invoked from the source host. It kicks off a series of message exchanges under unsupervised agent transport and eventually reaches the destination host. During the process, the source receptionist and destination receptionist are involved in the handshake. When the agent reaches the destination, it successfully unfreezes itself and is activated for normal execution. During the simulation, two indirect messages are sent to the agent owner/butler (pre-roaming and post-roaming notices) as stipu-

lated in the unsupervised agent transport protocol.

Functions to be carried out by the agents are loaded into the agent body before roaming. They will be preserved throughout agent transport. All functions carried are classified into sensitive functions and non-sensitive functions. Examples of sensitive functions are digital signature generation, negotiation strategy, and mission statement. Sensitive functions will be encrypted during the actual transmission. Non-sensitive functions refer to both functions with less sensitivity and functions that are vital to agent transport. Functions with less sensitivity do not need to be encrypted, and functions that are vital to agent transport cannot be encrypted; otherwise, it will not be able to perform regularly.

In the implementation, encryption on agent functions is done by first converting the agent function's byte-code into a binary stream (using the serialization feature of Java), and subsequently performing symmetric key encryption on the binary stream. The encrypted byte stream is carried in the agent body during agent transmission. When the agent reaches the destination, the encrypted byte stream will be decrypted into the original binary stream. From the original byte stream, the byte-code can be reconstructed and the agent function class can be dynamically loaded. The serialization feature of Java significantly reduces programming complexity here.

The flow of unsupervised agent transport protocol implementation is summarized below as an example:

1. Entry Permit Request (Agent to Source Receptionist)
Message content: agent certificate, destination address, and purpose of visit description.
2. Entry Permit Request (Source Receptionist to Destination Receptionist)
Message content: agent certificate, purpose of visit description.
3. Session Generation (Destination Receptionist)
Action: generate random session key, generate random challenge, generate freeze/unfreeze key pair, and store session information to local database.
4. Issue Entry Permit (Destination Receptionist to Source Receptionist)
Message content: agent description and entry permit (content of entry permit is discussed in the earlier section).
5. Entry Permit Reply (Source Receptionist to Agent)
Message content: entry permit.
6. Agent Freeze (Agent)
Action: generate random session key, encrypt sensitive functions with session key, encrypt session key with freeze key.
7. Pre-Roaming Notice (Agent to Agent Owner/Butler—Indirect)
The notice is sent as an e-mail message with destination address in the message body.
8. Send Request (Agent to Source Receptionist)
Message content: entry permit, destination address, and encrypted agent.
9. Send Agent (Source Receptionist to Destination Receptionist)
Message content: entry permit and encrypted agent.
10. Partial Activation (Destination Receptionist to Agent)
Action: activate the agent for execution to finish the agent transport process.
11. Unfreeze Key Request (Agent to Destination Receptionist)
Message content: agent certificate, entry permit, and session identifier.
12. Load Session (Destination Receptionist)
Action: validate entry permit, load un-

- freeze key from database based on session identifier.
13. Unfreeze Key Reply (Destination Receptionist to Agent)
Message content: unfreeze key.
 14. Unfreeze and Activation (Agent)
Action: decrypt session key using unfreeze key, decrypt sensitive functions using the session key.
 15. Post-Roaming Notice (Agent to Agent Owner/Butler—Indirect)
An e-mail is sent out to agent owner/butler notifying the success of agent transport.

CONCLUSION

SAFE is designed as a secure agent architecture for m-commerce. The foundation of SAFE is the agent transport protocol, which provides intelligent agents with roaming capability without compromising security. General security concerns as well as security concerns raised by agent transport have been carefully addressed. The design of the protocol also takes into consideration differing concerns for different applications. Instead of standardizing one transport protocol, three different transport protocols are designed, catering to various needs. Based on the level of control desired, one can choose between supervised agent transport and unsupervised agent transport. For applications that require a high level of security during agent roaming, bootstrap agent transport is provided so that individual applications can customize their transport protocols. The prototype of SAFE agent transport protocol has been developed and tested.

Agent transport protocol provides the secure roaming capability to SAFE. With a secure agent transport protocol, agents in SAFE can roam from host to host without being compromised. However, this does not complete the security framework in SAFE. Agent transport

protocol only addresses the security issues involved when the agent is roaming. There are other security issues to be considered. One of them is to protect agents against malicious hosts as well as protecting a host from malicious agents. To address these issues, agent flight recorder (AFR) and SAFE certification are being proposed. The design of AFR and SAFE certification will be studied in greater detail in the near future in order to complete the security framework for SAFE.

As an evolving effort to deliver a more complete architecture for agents, SAFER (Secure Agent Fabrication, Evolution, and Roaming) architecture is being proposed to extend the SAFE architecture. In SAFER, agents not only have roaming capability, but can make electronic payments and can evolve to perform better.

REFERENCES

- Bem, E. Z. (2000, December 11-15). Protecting mobile agents in a hostile environment. *Proceedings of the ICSC Symposia on Intelligent Systems and Applications (ISA 2000)*, Sydney, Australia.
- Corley, S. (1995, May). The application of intelligent and mobile agents to network and service management. *Proceedings of the 5th International Conference on Intelligence in Services and Networks (IS&N'98)*, Antwerp, Belgium.
- Finin, T., Fritzson, R., McKay, D., & McEntire, R. (1994). *KQML—A language protocol for knowledge and information exchange* (CS Technical Report CS-94-02). University of Maryland, USA.
- Finin, T., & Weber, J. (1993). *Draft specification of the KQML agent communication language*. Retrieved from <http://www.cs.umbc.edu/kqml/kqmlspec/spec.html>

- Gray, R. (1997). *Agent TCL: A flexible and secure mobile-agent system*. PhD Thesis, Department of Computer Science, Dartmouth College, USA.
- Guan, S. U., & Yang, Y. (1999). SAFE: Secure-roaming agent for e-commerce. *Proceedings of CIE '99*, Melbourne, Australia (pp. 33-37).
- Guilfoyle, C. (1994). *Intelligent agents: The new revolution in software*. London: OVUM.
- Johansen, D., Marzullo, K., & Lauvset, K. J. (1999, May 31-June 5). An approach towards an agent computing environment. *Proceedings of the ICDCS'99 Workshop on Middleware*, Austin, TX.
- Kotz, D., Gray, R., Nog, S., Rus, S., Chawla, S., & Cybenko, G. (1997). Agent TCL: Targeting the needs of mobile computers. *IEEE Internet Computing*, 1(4), 58-67.
- Odubiyi, J. B., Kocur, D. J., Weinstein, S. M., Wakim, N., Srivastava, S., Gokey, C., & Graham, J. (1997, February 5-8). SAIRE—A scalable agent-based information retrieval engine. *Proceedings of the Autonomous Agents 97 Conference*, Marina Del Rey, CA (pp. 292-299).
- Rus, D., Gray, R., & Kotz, D. (1996, August 4-5). Autonomous and adaptive agents that gather information. *Proceedings of the AAAI '96 International Workshop on Intelligent Adaptive Agents*, Portland, Oregon.
- Rus, D., Gray, R., & Kotz, D. (1997). Transportable information agents. In M. Huhns & M. Singh (Eds.), *Readings in agents*. San Francisco: Morgan Kaufmann.
- Sander, T., & Tschudin, C. F. (1998). Protecting mobile agents against malicious hosts. *Mobile Agents and Security/LNCS*, 1419, 44-60.
- Schneider, F. B. (1997, September 24-26). Towards fault-tolerant and secure agency. *Proceedings of the 11th International Workshop on Distributed Algorithms*, Saarbrücken, Germany.
- Schneier, B. (1996). *Applied cryptography: Protocols, algorithms, and source code in C* (2nd ed.). New York: John Wiley & Sons.
- Schoonderwoerd, R., Holland, O., & Bruten, J. (1997, February 5-8). Ant-like agents for load balancing in telecommunications networks. *Proceedings of the 1997 1st International Conference on Autonomous Agents*, Marina Del Rey, CA (pp. 209-216).
- Thirunavukkarasu, C., Finin, T., & Mayfield, J. (1995, December 1-2). Secret agents—A security architecture for the KQML agent communication language. *Proceedings of the CIKM'95 Intelligent Information Agents Workshop*, Baltimore.
- Westhoff, D. (2000, December 11-15). On securing a mobile agent's binary code. *Proceedings of the ICSC Symposia on Intelligent Systems and Applications (ISA'2000)*, Sydney, Australia.
- White, D. E. (1998). *A comparison of mobile agent migration mechanisms*. Senior Honors Thesis, Dartmouth College, USA.

Chapter XXVI

Tackling Counterfeiting with a Secure Online Track-and-Trace System

Fred Claret-Tournier
University of Sussex, UK

Pouwan Lei
University of Bradford, UK

Chris Chatwin
University of Sussex, UK

Rupert Young
University of Sussex, UK

ABSTRACT

This chapter introduces a secure online track-and-trace system for tackling counterfeiting. According to the Counterfeiting Intelligence Bureau (CIB), part of the International Chamber of Commerce, 7% of all world trade is in counterfeit goods and the counterfeit market is worth \$350 billion (IACC, 2003). Virtually every country in the world suffers from counterfeiting which results in—lost tax revenue, job losses, health and safety problems and business losses. Furthermore counterfeit goods do not only target famous brand names but anything that will sell such as bottled water. Counterfeiters are increasingly damaging businesses. Businesses need to fight against counterfeiting. Nowadays, there is an explosion of mobile wireless services accessible via mobile phones with a built in camera. The mobile users can access the Internet at any time, from anywhere using ubiquitous inexpensive computing. Mobile camera phones and other handheld devices are becoming indispensable. The aim of this chapter is to show how business can protect their products from counterfeiting by using a secure online track-and-trace system, which will allow their customers to authenticate the products in real time through a Web-enabled mobile camera phone. This will assist business and customers by confirming that the said product is genuine and not counterfeit, which can be accomplished at anytime and any location. Furthermore, the system does not require any significant change to the existing business operational systems.

INTRODUCTION

A counterfeit is something that is forged, copied, or imitated without the perpetrator having the right to do so, and with the purpose of deceiving or defrauding. Such rights are legally enshrined in patents, copyright, trademarks, industrial designs, and other forms of intellectual property protection (The Economist, 2003). A counterfeit product may be produced at one-tenth of the original cost of the genuine product with almost the same quality. Consequently, counterfeiting undermines fair competition, where manufacturers compete with each other on the basis of quality and price. Counterfeiting is a rapidly growing crime that is perpetrated on a global scale. There are a number of reasons for the widespread growth of counterfeiting crime:

1. There is a low risk of prosecution and enormous profit potential.
2. New technologies—especially information technology—enable counterfeiting in high volumes with high quality, for example, pirated music CDs, software, and movies. Pirated computer software and games account for half of the value of counterfeit products in the United Kingdom and costs the industry £3.29 billion (approx. \$5.9 billion), excluding the unpaid sales tax, each year (ACG, 2004).
3. The Internet is an effective and cheap communication channel for counterfeiters; it allows them to link consumers to suppliers with ease and relative anonymity in a global world. About \$25 billion worth of counterfeit goods are traded yearly over the Internet according to the Counterfeiting Intelligence Bureau (The Economist, 2003).
4. Countries opening up new markets such as China and Thailand produce more than 60% of the world's counterfeit goods in full-scale factories.
5. The advance in telecommunications, notably personal mobile phone systems, facilitates counterfeiters to be well informed by their lookout. In such a way, they can escape from law enforcement officers.

The objective of this chapter is to introduce a low-cost and secure online track-and-trace system for businesses to prevent counterfeiting by using a Web-enabled mobile phone. The problems of counterfeiting are discussed in detail, and the existing anti-counterfeiting technologies are reviewed. We introduce the architecture of the system, the area of application, and finally the conclusion.

Virtually every country in the world suffers from counterfeiting—lost tax revenue, job losses, health and safety concerns, and business losses. The economy is losing tax revenue and jobs because of the manufacture, distribution, and sale of counterfeit goods. In the United States, counterfeiting automobile parts, such as brake pads, cost the automotive industry alone over \$12 billion in lost sales; this is equivalent to hiring 200,000 additional workers. In total, American industry loses \$200-250 billion a year to counterfeiting activities. Since it is impossible to impose tax on counterfeit goods, governments lose tax revenue too. For instance, lost tax revenue in New York City alone total approximately \$350 million a year (IACC, 2003).

Counterfeiters are increasingly turning to mass-produced goods (EU, 2003). Counterfeited goods have expanded from luxury goods to consumer goods since these counterfeit products are not subject to authenticity checks. The European Commission's figures show that foodstuffs, cosmetics, toys, medicines, and car parts are prominently among the counterfeit goods seized. Due to the popularity of mobile phones, there is a marked increase in counterfeit mobile

phone products. The statistics show that seizures in 2002 were up almost 503% on 2000 and included many spare parts. Since these counterfeit products are not checked for conformity with safety standards, consumers are exposed to hazards.

The most worrying aspect of this crime is the widespread threat counterfeiting poses to public health and safety. The World Health Organization estimates that 5-7% of pharmaceuticals worldwide may be counterfeit—with too few active ingredients and many contaminants; fake labels or recycled packaging are used to cover up expiry dates (The Economists, 2003). The percentage of fake pharmaceuticals can rise to as high as 60% in developing countries. In the case of counterfeit medicines and cosmetics, it could endanger health as they may contain inactive or corrosive ingredients. According to a study in 1999, 10% of automotive spare parts sold in the EU are counterfeit. There is a flourishing trade in re-conditioned aircraft components, which are being sold as new and genuine parts, along with fake certificates of authentication. Counterfeit products can be found in almost every aspect of our lives—from alcohol to water, and pharmaceuticals to auto spare parts. It is damaging to our health, causes disasters, and destroys legitimate businesses.

Manufacturers, customers, and governments are the victims of this criminal activity. The manufacturer's brand is diluted and harmed by substandard goods damaging the reputation of the authentic product. The customer no longer has the reassurance of the genuine branded product. Tackling counterfeiting is an increasing problem as the range of counterfeit products widens considerably. The key to tackling counterfeiting is effective enforcement of brand protection schemes, which requires total commitment from business, stricter laws, and customer education. Counterfeit goods are of increasingly high quality so that it is becoming

difficult to distinguish them from a genuine product. To fight counterfeiting, business has to take the initiative to provide a mechanism for the stakeholders to verify if a product is genuine or not. At present famous brand name owners such as Gillette are making use of advanced anti-counterfeiting technologies to protect their products. There are a number of anti-counterfeiting technologies to protect and authenticate a genuine product. Some of the available technologies consist of: holograms, optical devices, smart cards, magnetic systems, printed labels, radio frequency identification (RFID), and smart molecule marking; these are discussed in the following section.

ANTI-COUNTERFEITING TECHNOLOGIES

Broadly speaking, there are two major types of anti-counterfeiting technologies: authentication technologies and track-and-trace technologies (FDA, 2003).

Authentication Technologies

These are categorized into the following groups:

1. **Overt** technologies are visible to the eye. They are holograms, moiré fringes, color shifting inks, and some watermarks. A hologram is an image encoded in phase and amplitude with the use of coherent laser light. The image is three dimensional, and can project deep inside or “stick” out of the hologram surface; it can be made virtually impossible to copy and display unique visual effects. When one grating, or a material with a periodic structure such as silk, slides over another moiré, fringes are produced. In making moiré measurements, two grids of evenly spaced lines are carefully prepared and super-

posed. One grid is never changed and is called the master grating. The other grating, affixed to a specimen under test, is called the specimen grating; this can be modulated to encode data for the unique labeling of goods.

2. **Covert** technologies are not visible to the eye and sometimes require special equipment for visualization (and authentication). These include some watermarks, certain inks and dyes that fluoresce or absorb ultraviolet light, and invisible bar codes. Watermark is an indelible image that is translucent (see-through) and embedded into the very fiber of the paper.
3. **Forensic technologies** require sophisticated analytical equipment. These include chemical markers, taggants, and other unique chemical properties of a substance. Using extremely small concentrations of chemicals, at the parts-per-billion levels, it is possible to mark products such as fuel without interfering with the quality of the product. The marker is detected and extracted from the product using its unique and specific recognition molecule. However, forensic marking usually requires expensive laboratory analysis in order to obtain a reliable result.

Overall, the authentication technologies have the advantage of being difficult or impossible to copy by any known technique. Most of them are machine-readable but require quite a long

time for verification. For instance, the analysis of a forensic mark has to be done in a highly sophisticated chemistry laboratory. There is a tradeoff between the duration of verification time and the degree of difficulty in copying.

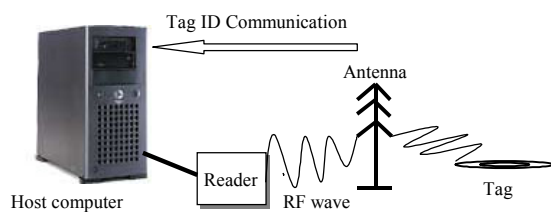
Track-and-Trace Technologies

These include auto-ID radio frequency identification (RFID) and barcodes. RFID is a form of contactless automatic identification and data capture (AIDC) technology for tracking and access applications, and employs a numbering scheme called EPC (electronic product code), which can provide a unique ID for any physical object in the world (Traiman, 2001). Each pack of cigarettes, individual can of soda, light bulb, or pack of razor blades produced can be made uniquely identifiable through its own EPC number which allows seamless, continuous identification and tracking of physical items as they move from one place to another, enabling companies to determine the whereabouts of all their products at all times (Schmidt, 2003). The ultimate goal of RFID is to create a “physically linked world” in which every item on the planet is numbered, identified, catalogued, and tracked (Shankar, 2001). An RFID system consists of the following components:

- antenna
- interrogator (reader)
- transponder (RFID tag)
- host computer

The tag and reader communicate with each other through an antenna over a radio frequency channel, which can travel easily through non-metallic materials (see Figure 1). Hence direct contact between the RFID tag and the reader is unnecessary. The reader, which may be fixed at a point or mobile (wireless or handheld) device, is linked to a host computer

Figure 1. The operation of RFID



through a local area network. Whenever the reader gets information from the tag, it can forward the data to the computer system and authenticate in real time. Also it enables multiple-read—that is, scanning a number of items simultaneously.

RFID tags come in a wide variety of shapes, sizes, and frequency ranges. Low-frequency systems have short reading ranges (less than 50 feet) and lower system costs. They are most commonly used in security access, asset tracking, and animal identification applications. High-frequency systems offer long read ranges (greater than 90 feet) and high reading speeds; these are used for such applications as railroad car tracking and automated toll collection. However, the higher performance of high-frequency RFID systems results in higher system costs. Under European law, the permitted frequencies and power for RFID radio signals are around eight times lower than those used in the United States (McCue, 2003). This means that the maximum accurate read range is around 1 foot.

A label contains information about product, supplier, customer, and carrier. Actually RFID is an intelligent label in which a reader reads a tag for information. The reader may write new or additional information to the tag which is somewhere between the size of a grain of sand and a speck of dust (Semilof, 2001). Furthermore, they can be built directly into food, clothes, drugs, or auto parts during the manufacturing process. RFID tags raise the problem of consumer privacy after the sale of a product with an embedded RFID tag, as they could be read without the consent of the tag carrier and would monitor individuals' behaviour. Retailers and manufacturers such as Wal-Mart, Gillette, Procter and Gamble, Tesco, and Coca Cola have attracted criticism from privacy groups (Albrecht & Founder, 2003).

For several decades, the bar code has been an indispensable tool for data collection that

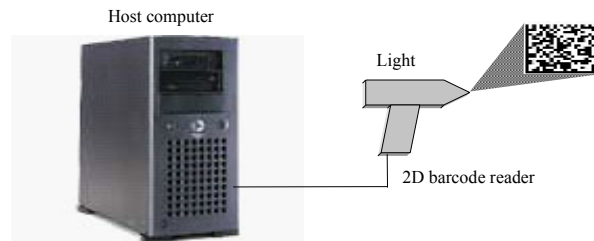
Figure 2. Matrix 2D and stacked 2D



supports many applications ranging from inventory tracking to point-of-sale identification. In today's market, every product has a barcode label on it, which serves to identify the product from a catalogue. There are several types of linear barcodes and also two-dimensional (2D) barcodes. The latter offers the advantage of encoding error-correcting codes in order to recover data from a partially damaged code (Poli, 1992; Peterson, 1972). The linear barcode is also called the *1D barcode*, a proven technology and probably the most reliable way for a machine to read a number. The main problem with the 1D barcode is that it does not include error-checking algorithms. Therefore the code is "vertically redundant," meaning that the same information is repeated vertically. The heights of the bars can be truncated without any loss of information. The vertical redundancy allows a symbol with printing defects, such as spots or voids, to still be read (Barnes, Bradshaw, Day, Schott, & Wilson, 1999). Hence, a 1D barcode encodes a small amount of data.

There are two distinct types of 2D barcodes—a stacked 2D barcode and a matrix 2D barcode (see Figure 2). Stacked 2D barcodes, such as PDF417 stacked bar codes, consist of thin slices of bar codes stacked on top of one another. Matrix 2D, for instance, data matrix, is composed of a matrix or grid arrangement of light and dark modules, and requires a 2D camera or image-based reader to be read and decoded (see Figure 3). Typically a data matrix can be as small as 5mm wide. However this size depends on the camera used for reading the code, the accuracy of the printer, and the amount of the data stored in the code. In

Figure 3. 2D barcode system



general, 2D matrix barcodes take up less than half the area of stacked 2D barcodes (AIM, 1996).

As far as the data matrix is concerned, the code is plotted as a series of black-and-white squares. The main advantage of data matrix, a 2D barcode, over the conventional 1D barcode is that it includes a very powerful error-correction algorithm called Reed Solomon. The error correction data stored in the symbol permits destroyed information to be mathematically recreated. Hence a badly damaged data matrix still recovers 100% of the information. In addition, it can hold between 100 and 2,000 characters in a single symbol because of compression techniques that can be applied; this reduces the physical size of the code and increases the amount of data that can be stored. The data printed in the code can also be encrypted to avoid easy counterfeiting. Printed labels are extremely cheap to produce, as they use existing technology and are easily mass-produced. Using laser technology, a data matrix can be marked permanently on products that are subject to extreme temperatures, aqueous wastes, and other chemical exposures. The downside is that they require a direct line of sight between the reader and the code, are read-only labels, and their content cannot be dynamically updated.

THE ARCHITECTURE OF THE SYSTEM

The system is designed for anti-counterfeiting where each product is labelled with a unique number. In order to avoid any possibility of copying the code, we use a random number generator to produce the unique number; this is encoded into a machine-readable object. The basic requirements of the system are traceability, shrinking surface area, speed, accuracy, and low cost. Amongst all the available anti-counterfeiting technologies that can be used to associate a code with a product, we have selected data matrix, due to its compact size, fault tolerance, encryption feature, and low cost. In addition, it will have little impact on the existing 1D barcode system, as a 2D barcode can incorporate the features of a 1D barcode. The information associated with the random number is stored in a secure online database that provides the product information such as product description, production location and date, and supply chain route. As the unique number code is read into the system, the online database is securely accessed using strong digital encryption. The data associated with the product code can be updated at any time during the lifecycle of the product, from the moment it is created until the moment it is sold.

Figure 4. Overview of online secure track-and-trace database system

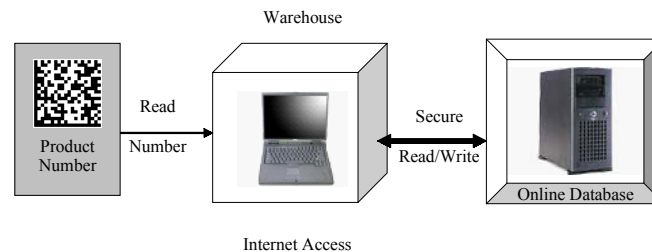


Figure 4 shows the architecture of the secure online track-and-trace system. The creation of the authentication code occurs when a product flows down the conveyor belt; at this time the unique code is registered to the database with the related product information. Furthermore, to speed up the scanning time by end users, the system also allows a number of products which are packed inside a box to be related to a group number, which is linked with all the unique product numbers inside the box. Each product has three numbers—a serial number, a unique product number, and a group number. The serial number represents the type of product (i.e., the information the 1D barcode has). The unique product number is the identification number, and the group number is the box number.

As the product is being sold, the buyer has two options when accessing the database: checking and registration. In checking, every stakeholder in the supply chain has the power to authenticate the product bought by scanning the product number and accessing the contents of the database. This scanning can be done at any stage within the supply chain, giving any relevant information about the product such as: the manufacturer's name, point of origin, destination, and manufacturing date. As a result, the buyer is made aware of the authenticity of the product. In registration, the stakeholder does not have to scan every product number,

but can use the unique group number. The system registers the name of the stakeholder and stores it in the database as the new owner of the products. This name would then be available for other distributors for product authentication.

As the product flows through the supply chain, the unique product number is scanned to authenticate the product, as well as giving a buyer proof of the sellers' ownership of the goods. The buyer is therefore aware of the product information stored in the database. The database is accessed through the Internet and gives a real-time response to the scanner, allowing the automation of the database uploading and downloading according to the privileges policy given to the user. The procedure to access and update the database is to scan the product number individually as the product changes ownership. It is essential to scan every item so that other users of the database can retrieve the correct data. From a practical point of view, only the box containing the said items needs to be scanned. The database software would automatically update every entry corresponding to the box serial number. On a larger scale, the transport container is coded in a similar way and the corresponding database entry is associated with every box code inside the container. Eventually, scanning the container's code updates the database entries for every item inside the container; the system

records the level of authentication. To prevent abuse of this approach, the external packaging requires a non-resealable seal.

The main goal of this system is to provide the stakeholders involved in manufacturing, distributing, selling, servicing, and buying with the ability to authenticate the product. In collaborating with each other with such a system, several advantages are realized:

1. The manufacturer will see the amount of counterfeit products being sold as genuine items reduce considerably. Its turnover could increase significantly, and profit levels would be increased.
2. The retailer benefits from the system by being able to prove that only genuine and authenticated items are being sold in his/her retail outlet. All the retailer needs to do is to update the database with the new owner of the product by scanning the label. The update notifies the database that the said product has been sold to an individual customer. On the other hand, the system can play a proactive role in preventing any stolen goods being used for a cash return. In this case the database would provide the retailer with the infor-

mation to authenticate whether the item had been sold or not.

3. Distributors along the supply chain would find it easier to sell authenticated products to other distributors with the authentication system in place, rather than products without any form of authentication. Note that in this case there is a major difference between authenticated products and authorized users/distributors. The aim of this system is to prove authenticity of a product and not canalize the distribution routes into authorized routes.

Decode Data Matrix with a Mobile Camera Phone

The commercial barcode reader is not a totally portable system, as it must be physically connected to a computer. Instead of using a barcode reader, we can make use of a portable mobile phone camera as an input device with which to capture the data matrix. The captured data matrix is decoded in the mobile phone. Then the decoded data is linked with a secure online database through the Web-enabled phone. In such a way a mobile user can access real-time information to authenticate the product he/she purchased. As for any portable device such as the smart phone or PDA, the amount of battery life and computing power available is critical for design of the application software. On a smart phone it is virtually impossible to apply standard image processing techniques to locate and decode the data matrix code. We developed specific image processing tools, adapted to the mobile phone environment, with an optimum design for this application. Figure 5 shows code reading and database access with a smart phone.

A 2D barcode reader uses an imaging system to grab an image of the barcode and extract data from that image. There are mainly two

Figure 5. Code reading and database access with a mobile phone

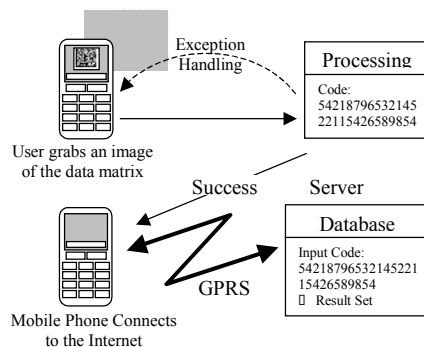


Table 1. Performance of data matrix symbol decoding

Task	Threshold	Finder Pattern	Density Bars	Binary Pattern	Error Correct	Total
Time (ms)	1-2	5-10	30	20	1-2	60

ways of designing software for image processing depending on whether the input scene is from a still image or from a live feed. In a still image scenario, only one image is available for processing, requiring powerful image processing techniques in order to achieve a positive recognition result. With a live image scene, an almost infinite number of still images are available, allowing the processing to be simpler and faster. At any point if an error is generated, a new image is automatically grabbed and the processing is started again. For a portable device such as the smart phone, the second approach is more suitable. Upon successful decoding of the code, a GPRS session is opened between the mobile phone and the global database for retrieval of product information.

Due to the limited amount of computational power of the mobile phone, pixel-based image processing operations are the most appropriate approach to the decoding of the data matrix. The processes are:

1. Establish quiet zone surrounding the symbol, where nothing other than the background color is printed.
2. Find pattern, which is the “L”-shaped solid lines usually situated on the left side and bottom side of the pattern.
3. Read the density bars, which are made of alternating dark and light modules.
4. Decode binary pattern, which contains the encoded data.

These four processes must be carried out in succession to achieve correct decoding. The image processing performed on the mobile phone needs to be optimized so that the duration of the processing remains within acceptable limits for real-time standards; therefore a few assumptions are made to avoid unnecessary use of computational power. The algorithm expects an input image that would comply with those assumptions in order to return a valid result. However, any error occurring during processing will result in the program grabbing another image. The image-processing algorithm was written in Java. This language has been chosen because of its portability, which means that the software can be uploaded to any phone supporting the Java language and providing that the C++ API interface with hardware modules such as the camera are supported.

Table 1 details the performances of the algorithm to decode a typical data matrix symbol within a 640x480 image. The performance also depends on the size of the data matrix symbol as well as the size of the input image. In this case, the program is able to process up to 16 images per second, using live image processing and decoding. The program is designed to act as a client-server interface, with the mobile phone as the client and the database as the server. The server is permanently listening to potential connections onto the database. When the client connects and sends a data matrix code in the required format, the database is accessed and information is retrieved and sent back to the mobile

Security Features

This system relies on creating a unique random number with no human interference over its randomness; the random number is then asso-

ciated with the relevant product data, which is held in a database. Security and granting privileges to database users are therefore major security issues. To ensure a secure communication between users and the online database, it is necessary to apply strong encryption in data transmission between any two parties. The type of encryption used is that of PKI (public key infrastructure) together with large public and private keys (Rivest, Shamir, & Adleman, 1978). This protocol remains one of the strongest encryption algorithms. The system administrator issues an individual public/private key pair to each system user, as well as a public/private key pair for himself. The system's administrator public key is stored in a certificate and the user's private key is stored in a file. Both of them are given to the user. The system's administrator keeps the certificate containing the public key associated with the user's private key. A certificate consists of public key, detailed information about the certificate owner (such as name, e-mail address), and other application-specific data (Recommendation X509, 1997).

The data sent between the user and the database is encrypted using the recipient public key. Only the receiver is able to decrypt the data. To authenticate the sender, two solutions are possible. The first solution is to encrypt the data first with the sender's private key and then with the recipient's public key. Then only the correct receiver can decrypt the data that only the correct sender could have had sent. The second solution is to use digital signatures, which are used to authenticate the originator of a message. Before the communication starts, the system generates a message digest of the message to be encrypted (Rivest, 1992). A message digest is the fixed-length result of a one-way hash of the contents in the message, similar to a cryptographic checksum or cyclic redundancy check (CRC). This message digest is encrypted using the recipient public key, and

the result is attached to the end of the message. After reception of the message, the receiver decrypts the message and the message digest, authenticating the sender and that the message has not been tampered with.

The database server must run on a secure operating system in order to avoid outside attacks. Limiting the number of registered users will simplify the security issue. Java is the chosen programming language in writing this application because of its capabilities to implement the strong encryption algorithms previously described and its automatic features such as memory reclamation of un-referenced memory, and the automatic checking of illegal array offsets. Furthermore, it does not allow a programmer to access the operating system and thus prevents any user from running malicious programs; this makes Java one of the safest programming languages, an important improvement over the C++ programming language.

The database serves not only for checking product details but also verifies the authenticity of the product. In checking, users may want to access the database to check the information related to a product such as date and physical location, if the product is in transit between warehouses. In some cases, users would need to have read and write privileges to the database. Since every stakeholder in the supply chain can access the online database, the access rights to users must be granted according to their relative importance in the supply chain. Some distributors would want to update the database with their personal details to reassure a potential buyer that the goods are using an authorized supply route. It is important for any business to authenticate transactions when dealing with, for instance, electronics or pharmaceuticals. It is not uncommon for counterfeit drugs to be imported into a country, or counterfeit electronic components to be assembled into products, without respecting the rules and regu-

lations of the country of destination. Using this system any company on the supply chain could check online and in real time whether the goods are the right goods for the right customers.

IMPLEMENTATION

In the following, we show how each stakeholder makes use of the secure online track-and-trace systems. A PCB assembly system is used as an example application. Besides having a serial number, every manufactured PCB is associated with a unique number, which is stored in an online database for any stakeholder to authenticate the goods. The unique number is printed on the label in the form of data matrix.

Manufacturer

A large amount of labels with unique numbers are pre-printed and transported to the manufacturer's production line where they are attached to the product. The registration and the creation of the database occur as the label is fixed onto the product. The set of labels is attached directly to individual products as they travel on the production conveyor belt. A fixed camera reads the labels and registers them online and in real time with the database (see Figure 6). Data associated with the product is then stored in the database. If the product is to

be stored in another container, a new code is attached to this new container and registered with the database. The container is sealed with a tamper-proof seal. The data associated with this container is the series of codes contained in it. The product is ready for shipping as soon as the label has been attached and the database entry is created, with the individual product codes associated with the global label.

Logistic Company

The PCB goods are transported from the manufacturer to the distributors using several possible combinations where the logistic company may belong to the manufacturer or the distributor or be a totally independent company. In either case the goods are scanned for authenticity and/or registered to the online database. The registration of the goods in the database allows product tracking and routing.

Distributor

The goods are scanned upon the receipt of the delivery. Since the logistic company has already done the registration, the distributor can check the authenticity of the goods and the route taken. Although this system represents an extra step in the stock management procedure, it allows the distributor's client to check the authenticity of the goods as well as the

Figure 6. Registering codes on the production line

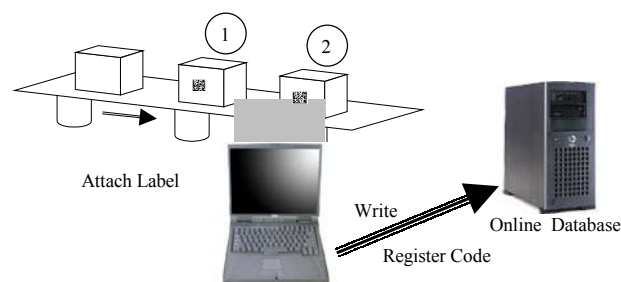
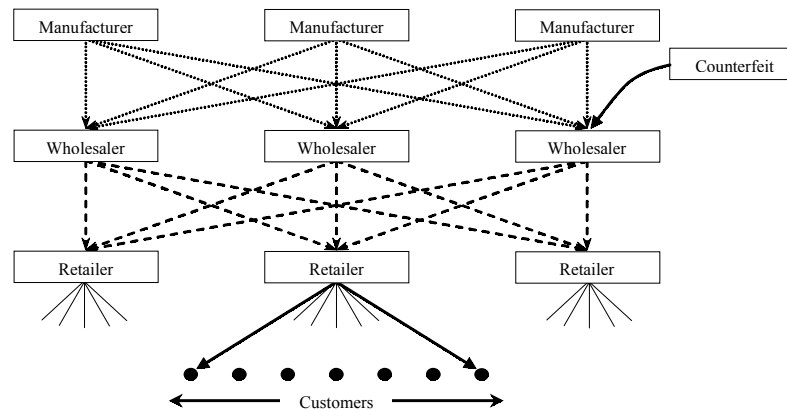


Figure 7. The Retailer can check the authenticity of products in the supply chain



legality of the goods. That is, do they belong to the distributor or not? It would be impossible for the client not to be aware of the introduction of counterfeit products or distribution via unauthorized channels.

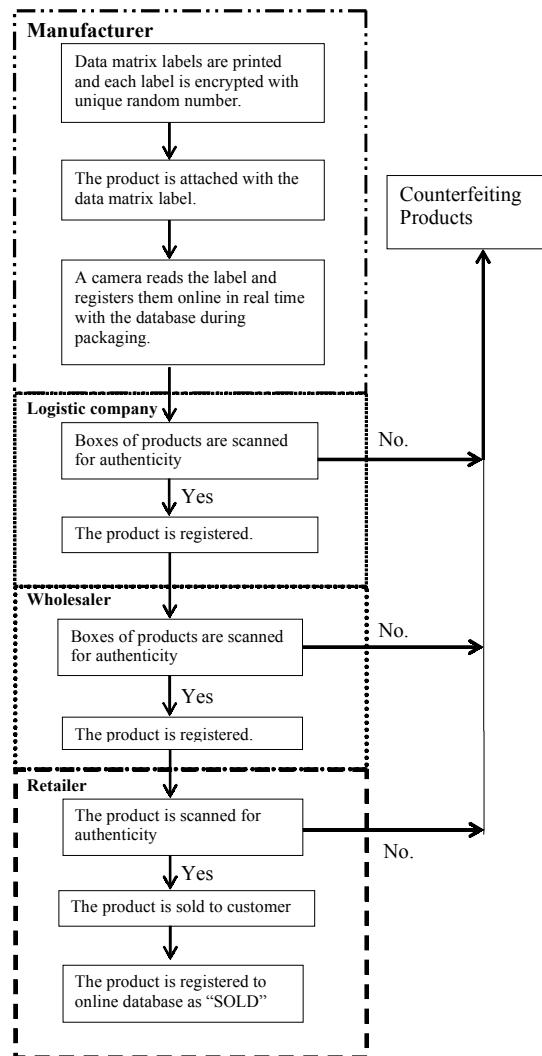
Retailer

Being the end-of-line stakeholder, the retailer is in direct contact with the customer. By using this system, the retailer could guarantee customers that only genuine goods are available in his retail outlet and that no counterfeit goods could infiltrate his distribution network. The retailer could scan the product/group code on reception of the goods for authenticity (see Figure 7). At the same time, the retailer can make use of this system to deal with false returns when stolen goods are used to claim a refund. If the product is sold, the database is updated with a status such as “sold.” When the product is returned, the status of the product can be checked from the database. If the product was actually sold, it can then be returned. On the other hand, if the status is “not sold,” the item would not be refunded. Figure 8 shows the processes in the system.

CONCLUSION

The system provides users with secure access to the online database where information about the product is stored. This database access is achieved using a Web-enabled mobile camera phone. The client/server software uses the most secure and reliable encryption and certificate standards for data communication. Since the amount of data transmitted between the server and client is small, there is a fast response in retrieving data from the database. The system can therefore be used at any point in the supply chain. The unique numbering of products prevents counterfeit goods from being introduced into the supply chain. With the system, the retailers will be able to authenticate products. The scanning process is easy and relatively inexpensive. Other technologies such as RFID could be used in conjunction with the printed barcode. For instance, RFID can be used for the unique group number on the outside of a box or a container. An optimum compromise between system cost and practicality will always determine the exact nature of system deployment.

Figure 8. The system's processes



Furthermore, the implementation of the system does not require any significant change to the existing business operational system. A Web-enabled mobile camera phone will decode the 2D barcode label. This makes possible the reading of 2D barcodes at anytime and in any location. The main goal of this system is to provide the stakeholders involved in manufacturing, distributing, selling, servicing, and buying with the ability to authenticate the product.

In collaborating with each other with such a system, several advantages are realized:

1. The manufacturer will see the amount of counterfeit products being sold as genuine items reduce considerably. Its turnover could increase significantly and profit levels would be increased.
2. The retailer benefits from the system by being able to prove that only genuine and authenticated items are being sold in his/her retail outlet. All the retailer needs to do is to update the database with the new owner of the product by scanning the label. The update notifies the database that the said product has been sold to an individual customer. On the other hand, the system can play a proactive role in preventing any stolen goods being used for a cash return. In this case the database would provide the retailer with the information to authenticate whether the item had been sold or not.
3. Distributors along the supply chain would find it easier to sell authenticated products to other distributors with the authentication system in place, rather than products without any form of authentication. Note that in this case, there is a major difference between authenticated products and authorized users/distributors. The aim of this system is to prove authenticity of a product and not canalize the distribution routes into authorized routes.
4. Consumers have trust in the product they purchased and the business's reputation is preserved.

This cost-effective and secure mobile track-and-trace system will enable businesses to fight counterfeiting crime, which destroys stakeholder business confidence. For long-term business success, businesses should behave ethi-

cally and be responsible to their stakeholders; this technology significantly improves their ability to achieve this.

REFERENCES

- ACG (Anti-Counterfeiting Group). (2004). Retrieved April 8, 2005, from http://www.acg.com/guest_frames04.html
- AIM. (1996). *International symbology specification—data matrix*. USA: Association for Automatic Identification and Mobility Global.
- Albrecht, K., & Founder, C. (2003). *RFID: Tracking everything, everywhere*. Retrieved April 8, 2005, from http://www.boycottgillette.com/rfid_overview.html
- Barnes, D., Bradshaw, J., Day, L., Schott, T., & Wilson, R. (1999). *Two-dimensional bar coding*. Tech 621, Purdue University, USA.
- The Economist. (2003, May 17-23). Special report: Counterfeiting: Imitating property is theft. *The Economist*, p. 69-71.
- EU (European Commission). (2003, October 24). *Customs: Counterfeiters and pirates are increasingly turning to mass-produced goods*. Retrieved April 8, 2005, from <http://www.cemarking.net/article/articleview/376/1/88/>
- FDA. (2003). *Counterfeit Drug Task Force interim report*. Rockville, MD: U.S. Food and Drug Administration.
- IACC (International AntiCounterfeiting Coalition). (2003). *Facts on fakes*. Retrieved from www.iacc.org
- McCue, A. (2003, October 16). UK retailer tests radio ID tags. *CNET News.com*. Retrieved April 8, 2005, from <http://news.com.com/2100-1039-5092460.html>
- Peterson, W. W., & Weldon, E. J. (1972). *Error-correcting codes*. Cambridge, MA: MIT.
- Poli, A. (1992). *Error correcting codes: Theory and applications*. Hemel Hempstead, UK: Prentice-Hall.
- Recommendation X509. (1997). *Information technology—Open systems interconnection—The directory: Authentication framework*. Geneva, Switzerland: The International Telecommunications Union.
- Rivest, R. (1992). *The MD5 Message Digest Algorithm [RFC 1321]*. Engineering Task Force. Retrieved February 12, 2006, from <http://www.faqs.org/rfcs/rfc1321.html>
- Rivest, R., Shamir, A., & Adleman, L. (1978). A method for obtaining digital signature and public-key cryptosystems. *Communications of the ACM*, 21(2), 120-126.
- Schmidt, C. W. (2003). The networked physical world. *Beyond the Internet*, Retrieved April 8, 2005, from http://www.rand.org/scitech/stpi/ourfuture/Internet/sec4_networked.html
- Semilof, M. (2001, November 19). Bar codes in a chip. *InternetWeek.com*. Retrieved April 8, 2005, from <http://www.internetweek.com/newslead01/lead111901>
- Shankar, M. K. (2001). *Algorithm ensures unique object ID*. Retrieved April 8, 2005, from http://www.nikkeibp.asiabiztech.com/nea/200104/inet_127161.html
- Traiman, S. (2001). *Tag, you're it! The ePC tag could revolutionize the retail supply chain*. Retrieved April 8, 2005, from http://www.retailsystemsreseller.com/archive/Nov01/Nov01_5.shtml

Chapter XXVII

Developing a Theory of Portable Public Key Infrastructure (PORTABLEPKI) for Mobile Business Security

Sashi Nand

Rushmore University, Grand Cayman, BWI

ABSTRACT

The issue of security is paramount for the success of mobile business. Although the state of wirelessness offers portability, and therefore mobility, it adds to the risk of unauthorised access to the system and data disclosure. This chapter discusses how public key infrastructure (PKI) technology can be implemented to reduce the risks associated with mobile business. A theory of portable PKI (PORTABLEPKI) developed in this chapter within the context of Australian industries is to promote PKI technology for enhancing the security of mobile business. A framework for testing PORTABLEPKI theory and future research opportunities which will open up as a result of this developmental study are also provided.

INTRODUCTION

This chapter looks at how a public key infrastructure (PKI) can increase the wireless network's security by requiring certificate-based authentication for access. It also develops a theory of PORTABLEPKI. Finally, a framework for testing PORTABLEPKI and future research opportunities are discussed.

MOBILE BUSINESS

Mobile Business (m-business) can simplistically be understood as follows:

M-Business = Internet + E-Business + Wireless

M-business is the application infrastructure required to maintain business relationships by

means of mobile devices. M-business is also the logical extension of electronic business (e-business) to address new customer channels and integration challenges. There is an inter-connection of business processes within an organization and between external parties. For the notion of “business without boundaries” to prevail, back-end applications and data must be re-engineered to take complete advantage of the features offered by m-business (Kalakota & Robinson, 2002).

The most challenging and complex aspects of the m-business revolution are the design implementation, security, and integrity of mobile-enhanced business processes because they transcend traditional and regulatory boundaries (Stanley, 2004).

WIRELESS NETWORK

Wireless technologies are based on communication without land-based physical connections. For example, traditional telephone handsets use continuous cabling for connectivity, hence it is wired. Wireless telephony, on the other hand, uses radio waves rather than cables to broadcast network traffic and data transmission.

The two primary areas of wireless technology are mobile phones and mobile computers. Mobile implies portability—a device such as a mobile phone, PalmPilot, or laptop that travels with the user and can be used either off-line or online:

- *Mobile and off-line* means that the device can be used to run self-contained applications while not connected to the Internet or other telephony devices.
- *Mobile and online* is commonly called wireless. This means that the experience is based on a live connection supplied via satellite, cellular, or radio transmission.

An online device will always be ‘on’ in the presence of any wireless network—seamlessly connecting to the Internet or some other system (Kalakota & Robinson, 2002).

What is a Wireless Network?

In a wireless network, radio waves carry the signal at least part of the way. The greater the proportion of the wireless to wired, the more wireless we consider the network. Three basic wireless networking technologies include:

- **Wireless Private Area Networks (WPANs):** Refer to confined short-range networks, for example computers connected while traveling such as mobile phones, laptops, and personal digital assistants (PDAs).
- **Wireless Local Area Networks (WLANs):** Refer to same local-range networks, for example computers connected within the same area such as an office building or home.
- **Wireless Wide Area Networks (WWANs):** Refer to long-range networks, for example computers connected over long distances such as a university campus, city, or town (Shaw, 2003).

SECURITY

With any new technology—especially wireless networking—concerns and questions arise about security of data transmission (Shaw, 2003). Security is a process of minimizing risk, threat, or the likelihood of harm (Pipkin, 2000).

Wireless communications are inherently more open to attack than wired data transfer because the physical layer is the uncontained cyberspace (Campbell, Calvert, & Boswell, 2003).

An insecure wireless connection exposes users to intrusion, which can lead to a loss of protection for confidential information, interception of messages, or abused connections. Some examples are:

- E-mail can be intercepted, read, or changed.
- A hacker who hijacks a session can replace a user's credentials with false information gaining access to the system.
- An unauthorized person can log on to a wireless network that is not secure and use the resources, or obtain financial gain through deception including free connectivity to the Internet (Chan, 2004).

Security dominates discussions about wireless communication. The reason is simple: removing the wires simultaneously removes the access restrictions. In fact, many wireless networks begin life completely unsecured because vendors design wireless access points (WAPs) and WLAN cards with ease of installation and usage in mind. Configuration of security settings does not equate with ease of use. For this reason, a secure network needs to be set up intentionally and consciously (Randall & Sosinsky, 2005).

Even the most technically efficient and well-managed wireless network will be of little use if the network is not secured (Shaw, 2003). When implementing a wireless network, a plan must be developed for securing the network to reduce the likelihood of risks and threats.

METHODS OF SECURING WIRELESS NETWORKS

Security of wireless networks is specifically covered by the Institute of Electrical and Electronic Engineers (IEEE) 802.11i security speci-

fication. Some of the common methods of protecting a wireless network are as follows.

Media Access Control (MAC) Filtering

One of the most basic ways of protecting a wireless network is to implement MAC filtering. At the WAP, configure those MAC addresses (the low-level firmware address of a wireless card) that are allowed to connect to the WAP. Although this sounds like an ideal and easy way to secure a wireless network, consider the following weaknesses:

- It is easy to spoof an approved MAC address.
- MAC filtering is hard to manage.
- MAC filtering authenticates only the computer, not the user.
- The size of the approved MAC list is limited.

Wired Equivalent Privacy (WEP)

WEP provides encryption services to wireless networking. When a wireless connection enables WEP, the wireless network interface card (NIC) encrypts each data packet transmitted on the network using the Rivest Cipher version 4 (RC4) stream cipher algorithm. The WEP then decrypts the data packets on receipt. The weakness in WEP's implementation is two-fold:

- The symmetric encryption key is rarely changed.
- The initialization vector (IV) is only 24 bits and is re-used over time, thereby giving rise to a pattern of usage which can be easily identified and exploited (Komar & Microsoft PKI Team, 2004).

Wi-Fi Protected Access (WPA)

This is an encryption standard produced by the Wireless Fidelity (Wi-Fi) Alliance to address the security issues found in WEP. The following enhancements are included in WPA:

- **Increased Data Encryption:** WPA implements Temporal Key Integrity Protocol (TKIP), which uses a per-packet key mixing function, a message integrity check (MIC) known as *Michael*, and an extended IV with rules on sequencing. In addition, WPA implements a re-keying mechanism so that the same key is not used for long periods of time.
- **Dependency on 802.1x Authentication:** The use of 802.1x authentication is optional for WEP encryption only. WPA requires 802.1x authentication to ensure that only authorized users or computers are allowed to connect to the wireless network. 802.1x authentication also ensures mutual authentication so that a wireless client does not connect to a rogue network, rather than an authorized network.

Weaknesses in the current WEP algorithm implemented in current WLANs have been exposed. The new security supplement to the 802.11 MAC standard is 802.11i, which will address security holes in the 802.11a, b, and g protocols, and improve encryption, key management, distribution, and user authentication. This standard is worth remembering, because these improvements to security may be available as firmware and later hardware upgrades for existing Wi-Fi networks (McCullough, 2004).

The current WPA definition includes forward compatibility with the new 802.11i security specification. 802.11i adds secure fast handoffs, secure de-authentication, and secure

disassociation with WAPs. 802.11i also implements strong forms of authentication from the Advanced Encryption Standard (AES) (Komar & Microsoft PKI Team, 2004).

Public Key Infrastructure

PKI significantly increases the security of wireless networks because it requires encryption as well as a certificate-based authentication for access. PKI uses pairs of cryptographic keys (public key and private key) provided by a trusted third party, known as a certification authority (CA), which is verified by a registration authority (RA). Central to the workings of PKI, a CA issues a digital certificate, which positively identifies a holder of keys. The CA maintains accessible directories of valid certificates and also a list of certificates it has revoked.

PKI brings to the electronic world the security and confidentiality normally provided by physical documents such as handwritten signatures, sealed envelopes, and established trust relationships that are part of traditional paper-based transactions. These security and confidentiality features are as follows:

- **Confidentiality:** Ensures that only intended recipients can read files, or changes can only be implemented with a valid key.
- **Data Integrity:** Ensures files cannot be changed.
- **Authentication:** Ensures that participants in an electronic transaction are who they claim to be.
- **Non-Repudiation:** Prevents participants from denying involvement in an electronic transaction (Austin, 2001).

Vendors can provide security solutions that install digital certificates on the end devices itself, optimizing the PKI implementation espe-

cially for the wireless environment. For example, Microsoft Windows Server 2003 PKI provides the necessary certificates for 802.x authentication for wireless as well as wired networks. When a user or computer performs 802.1x authentication for wireless or wired network, the following two authentication types are available:

- **Extensible Authentication Protocol with Transport Layer Security (EAP/TLS):** A certificate-based authentication method that provides mutual authentication between the user or computer and the Remote Authentication Dial-In User Service (RADIUS) server when implemented for a wireless networking solution.
- **Protected Extensible Authentication Protocol (PEAP):** Allows the transmission of other EAP types within a TLS secure channel (Komar & Microsoft PKI Team, 2004).

Thus, in an open, untrusted, and insecure wireless network environment, cryptography provides the security and PKI provides the trust to enhance m-business (Deloitte & Touche Research Team, 2001).

The lack of security in mobile business is the fundamental problem. In order to better understand and examine ways in which this problem may be solved, a research theory that addresses this critical issue of security using PKI is developed in the following section.

DEVELOPING A THEORY OF PORTABLE PKI (PORTABLEPKI)

There can be situations where it is absolutely critical to use PKI, but it comes at a cost. Nevertheless, despite its costs and complexity, laying down a sound theoretical foundation,

combined with best business practices and robust technological infrastructure, will enhance the usage of PKI and the security of mobile business. Hence Portable Theory of PKI (PORTABLEPKI) has been developed.

There is no definite meaning given to the term “theory,” and there are many views on what constitutes a theory. The standard or the orthodox view has been used to construct PORTABLEPKI theory. According to this view there exists a phenomena in the real world (P-Field). Observation of phenomena leads to abstractions by an individual’s reason (C-Field) (Staunton, 1976). A theory begins in the ‘unreal’ world of abstraction, that is, in the human mind (C-Field). In order for it to be useful, theory must eventually relate to the ‘real’ world, the world of experience (P-Field). Three types of relationships in the theoretical structure are:

1. **Syntactics:** Rules of language. If expressed in English, then the relationship refers to the rules of grammar. If the theory is mathematical, then the relationship refers to the rules of mathematics.
2. **Semantics:** Rules of correspondence or operational definitions which link the concepts to objects in the real world. Semantics concern the relationship of a word, sign, or symbol to a real-world object or event. It is the semantic relationship that makes a theory realistic and meaningful.
3. **Pragmatics:** The effect of words or symbols on people. We are interested in how concepts and their measured correlations in the real world affect people’s behavior (Nand & Unhelkar, 2003).

The first step of PORTABLEPKI theory is to identify the research problem in the P-Field by observing the use of PKI in the real world of mobile business. The next step is to develop the conceptual and theoretical structure, including

the causal links and chains, and state the hypothesis (H). Then the hypothesis can be written in simple English stating the relationship of each clause whether they are directly or indirectly related. The real-world effect on stakeholders also has to be shown. The overall theory of PORTABLEPKI can be viewed as a set of principles for the purpose of enhancing growth and acceptance of PKI, as well as to enhance the security of m-business. A framework for testing this theory is provided in the next section.

A FRAMEWORK FOR TESTING PORTABLEPKI

An empirical research program based on the inductive-deductive approach developed by Abdel-Khalik and Ajinkya can be modified to test this theory of PORTABLEPKI (Godfrey, Hodgson, & Holmes, 1997). This involves the following eight stages:

Stage 1: Identify a Research Problem by Observation

PKI is one of the remedies to m-business security problems. An examination can be made of the following six factors influencing use of PKI technology in m-business in Australia (Nand & Unhelkar, 2003):

- **Environment:** Includes security, globalization, market competition, regulating forces, telecommunications, and political influence.
- **Organization:** Includes corporate governance, management, organizational structure, and resources.
- **Business Strategy:** Includes strategic planning, business process re-engineer-

ing, total cost of ownership, and return on investment.

- **IT Strategy:** Includes strategic planning, system development, system maintenance, technological risk (including wireless), and complexity of PKI.
- **PKI Technology:** Includes necessity of trust, PKI initiatives, PKI availability, and PKI success stories.
- **People:** Includes PKI skills, PKI training and dissemination of information, and employee culture.

Stage 2: Develop the Conceptual and Theoretical Structure, Including Causal Links and Chains

To develop a rationale as to why firms do or do not use PKI, the study would test the effect of some selected independent variables for the use of PKI. Two independent variables which influence use of PKI are industry type (service or non-service) and the number of years of IT experience. The dependent variable is the level of usage of PKI.

Stage 3: Operationalize the Theoretical Constructs and Relationships, and State the Specific Hypothesis to be Tested

Two hypotheses that have been developed to test this theory are:

- **Hypothesis 1:** Higher usage of PKI technology is expected in the service industry compared with the non-service industry.
- **Hypothesis 2:** Greater usage of PKI technology is expected in organizations that have a greater number of years of IT experience.

Stage 4: Construct the Research Design

The survey research method can be adopted to obtain data from organizations Australia-wide.

Stage 5: Implement this Design by Sampling and Gathering Data

A sample of Australian companies from at least one service industry and one non-service industry can be selected and company details recorded using a database.

Stage 6: Analyze Observations in Order to Test Each Hypothesis

Descriptive statistics and Chi-Square Test can be used to process and analyze the collected data using Microsoft Excel together with PHStat, Prentice-Hall's statistical add-in for Excel. With descriptive statistics frequency, distributions of all responses to the national survey can be recorded using simple tabulations and cross-tabulations on the Microsoft Excel spreadsheet. By using Chi-Square Test, hypotheses 1 and 2 can be tested together with PHStat. This test involves comparison of actual frequency with expected frequency.

Stage 7: Evaluate the Results

Determine whether or not the results support the theory of PORTABLEPKI.

Stage 8: Consider the Specific Limitations and Constraints

Refer to the procedures undertaken in Stages 1-7, and ask: Are there any limitations to the way the theory was developed or tested? Do any refinements to the theory appear warranted? If the answer is 'yes' to either question,

then return to the appropriate stage and attempt to remedy the limitation.

CONCLUSION AND FUTURE DIRECTIONS

Theories play an important role in understanding and changing the world. This chapter has developed a theory of PORTABLEPKI, which breaks new ground. The next step is to test the hypotheses stated in this chapter using the framework provided here and to validate the reality of PKI usages. This will lead to the refinement of PORTABLEPKI theory. People dealing with information security systems have to be ever vigilant because security is an unending mission. While creativity and innovation are what drives new technology, it also gives rise to its associated security problems. Hence the implementation of PORTABLEPKI theory will lead to increased usage of PKI which consequently will enhance the security of both wired and wireless networks and mobile businesses.

This is an initial step for the development of the theory of PORTABLEPKI for the purpose of increasing the security of m-business. The future direction is for the PORTABLEPKI theory to be tested by selecting one specific service industry (e.g., automobile telematics (wireless telemetry) industry) and one non-service industry (e.g., mining oil and gas). This testing will either confirm or negate the PORTABLEPKI theory. Further work could then be undertaken to refine this theory for other stages of development in m-business.

REFERENCES

Austin, T. (2001). *PKI: A Wiley tech brief*. New York: John Wiley & Sons.

- Campbell, P., Calvert, B., & Boswell, S. (2003). *Security + guide to network security fundamentals*. Boston: Cisco Learning Institute, Thomson Course Technology.
- Chan, D. (2004). What auditors should know about encryption. *Information Systems Control Journal*, 3, 32.
- Deloitte & Touche Research Team. (2001). *E-commerce security: Public key infrastructure: Good practices for secure communications*. Rolling Meadows, IL: Information Systems Audit and Control Foundation.
- Godfrey, G., Hodgson, A., & Holmes, S. (1997). *Accounting theory* (3rd ed.). Sydney: John Wiley & Sons.
- Kalakota, R., & Robinson, M. (2002). *M-business: The race to mobility* (pp. 8-10, 19). New York: McGraw-Hill.
- Komar, B., & Microsoft PKI Team. (2004). *Microsoft Windows Server 2003 PKI and certificate security* (pp. 467-471). Redmond, WA: Microsoft Press.
- McCullough, J. (2004). *185 wireless secrets: Unleash the power of PDAs, cell phones, and wireless networks*. Indianapolis: Wiley Publishing.
- Nand, S., & Unhelkar, B. (2003, November 24). Progress report on development of "Investigations Theory of PKI" and its application to Australian information systems. *Proceedings of the 1st Australian Information Security Management Conference*, Perth, Australia (p. 3).
- Nand, S., & Unhelkar, B. (2003, December 16-18). Development of an Australian trust scheme of PKI to enhance confidence in security for e-transforming organisations: A study of a cluster of SMEs in Australia. *Proceedings of the 2003 International Business Information Management Conference*, Cairo, Egypt (p. 5).
- Pipkin, D. L. (2000). *Information security: Protecting the global enterprise*. Upper Saddle River, NJ: Prentice-Hall.
- Randall, N., & Sosinsky, B. (2005). *PC Magazine: Wireless solutions*. Indianapolis: Wiley Publishing.
- Shaw, R. (2003). *Wireless networking made easy*. New York: AMACOM.
- Stanley, R. A. (2004). Security, audit and control issues for managing risk in the wireless LAN environment. *Information Systems Control Journal*, 3, 23.
- Staunton, J. J. (Ed.). (1976). *Theory construction and verification in accounting*. Armidale, Australia: University of New England.

Chapter XXVIII

Systems, Handheld Devices, and Payment Methods for Mobile Commerce

Wen-Chen Hu

University of North Dakota, USA

Tom Wiggen

University of North Dakota, USA

Hung-Jen Yang

National Kaohsiung Normal University, Taiwan

ABSTRACT

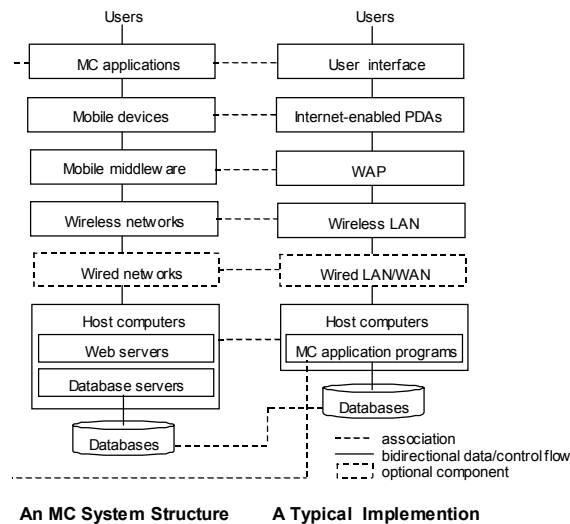
The emergence of wireless and mobile networks has made possible the introduction of electronic commerce to a new application and research subject: mobile commerce. Understanding or constructing a mobile commerce system is an arduous task because the system involves a wide variety of disciplines and technologies. This chapter tries to relieve this problem by giving careful studies of three themes of mobile commerce: (1) mobile commerce systems and transactions of which a system includes six components: (a) mobile commerce applications, (b) mobile handheld devices, (c) mobile middleware, (d) wireless networks, (e) wired networks, and (f) host computers; (2) mobile handheld devices, which are the communication devices between mobile applications and users and also include six major components: (a) a mobile operating system, (b) a mobile central processor unit, (c) a microbrowser, (d) input/output devices, (e) a memory, and (f) batteries; and (3) mobile payment methods, which include macro- and micro-payment methods. Other important issues, such as mobile commerce transactions and mobile security properties, are also discussed.

INTRODUCTION

The introduction of the World Wide Web has revolutionized traditional commerce and boosted

sales and exchanges of merchandise and information. Recently, the emergence of wireless and mobile networks has made possible the extension of electronic commerce to a new

Figure 1. A mobile commerce system structure



application and research area: mobile commerce. This technology allows the exchange or buying and selling of commodities, services, or information on the Internet through the use of mobile handheld devices. In just a few years, mobile commerce has become the most demanding trend in business transactions. Despite a weak economy, the future of mobile commerce is bright according to the latest predictions (Reuters, 2001). However, it requires a tremendous effort to understand mobile commerce and construct a mobile commerce application, because mobile commerce involves such a wide range of disciplines and technologies. To explain the problem, this chapter carefully studies three major subjects of mobile commerce:

1. **Mobile Commerce Systems:** Various system structures for mobile commerce have been proposed. A typical mobile commerce system generally consists of six components: (1) mobile commerce applications, (2) mobile handheld devices, (3) mobile middleware, (4) wireless net-

works, (5) wired networks, and (6) host computers.

2. **Internet-Enabled Mobile Handheld Devices:** Handheld devices provide the equipment mobile users need to interact with mobile commerce applications. A mobile handheld device includes six major components: (1) a mobile operating system, (2) a mobile central processor unit, (3) a micro browser, (4) input/output devices, (5) a memory, and (6) batteries.
3. **Mobile Payment Methods:** Mobile commerce security is defined as the technological and managerial procedures applied to mobile commerce to provide security properties. Among the many issues that arise with mobile commerce security, mobile payment methods, including macro-payments and micro-payments, are probably the most important.

BACKGROUND

This section provides the necessary background information of the three themes of this chapter: mobile commerce systems, Internet-enabled mobile handheld devices, and mobile commerce payment methods.

Mobile Commerce Systems

A mobile commerce system is inherently interdisciplinary and could be implemented in various ways. Figure 1 shows the structure of a mobile commerce system and a typical example of such a system (Hu, Lee, & Yeh, 2003). The system structure includes six components:

1. mobile commerce applications,
2. Internet mobile handheld devices,
3. mobile middleware,
4. wireless networks,

Figure 2. A flowchart of a user request processed in a mobile commerce system

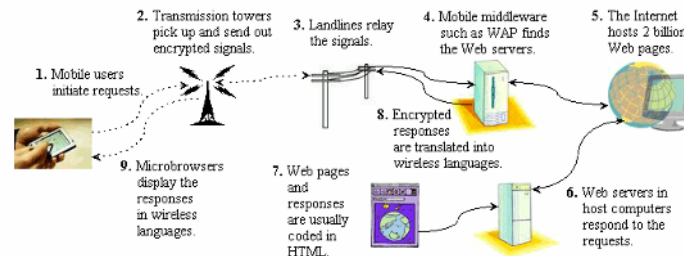
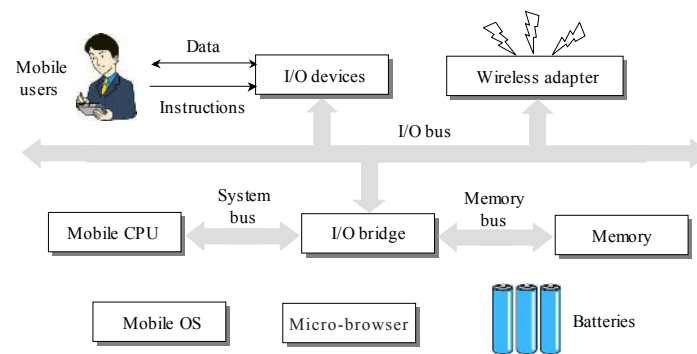


Figure 3. System structure of mobile handheld devices



5. wired networks, and
6. host computers.

To explain how the mobile commerce components work together, Figure 2 shows a flowchart of how a user request is processed by the components in a mobile commerce system.

Internet-Enabled Mobile Handheld Devices

Mobile users interact with mobile commerce applications by using small wireless Internet-enabled devices, which come with several aliases such as handhelds, palms, PDAs, pocket PCs, and smart phones. To avoid any ambiguity, a general term, mobile handheld devices, is

used in this chapter. Mobile handheld devices are small general-purpose, programmable, battery-powered computers, but they are different from desktop PCs or notebooks due to the following special features:

- limited network bandwidth,
- small screen/body size, and
- mobility.

Figure 3 shows a typical system structure for handheld devices, which includes the following six major components: (1) a mobile operating system, (2) a mobile central processing unit, (3) a micro-browser, (4) input/output devices, (5) a memory, and (6) batteries.

Mobile Commerce Payment Methods

Mobile commerce security is defined as the technological and managerial procedures applied to mobile commerce to provide security properties, which will be explained next. Among the many issues that arise with mobile commerce security, mobile payment methods are probably the most important. They are the methods used to pay for goods or services with a mobile handheld device, such as a smart cellular phone or an Internet-enabled PDA. There are usually two kinds of mobile commerce payment methods:

- **Macro-Payments:** These kinds of payments are used by traditional electronic commerce and usually involve amounts more than US\$10.
- **Micro-Payments:** These usually involve amounts less than US\$10, which is too small to be economically processed by credit cards.

MOBILE COMMERCE SYSTEMS

Mobile commerce is an effective and convenient way to deliver electronic commerce to consumers from *anywhere* and at *anytime*.

Realizing the advantages of mobile commerce, many major companies have begun to offer mobile commerce options for their customers in addition to the electronic commerce already provided (The Yankee Group, 2001). However, it requires a tremendous effort to understand or construct a mobile commerce system because it involves such a wide range of disciplines and technologies. To lessen the difficulty, this chapter will divide a mobile commerce system into six components: (1) mobile commerce applications, (2) mobile stations, (3) mobile middleware, (4) wireless networks, (5) wired networks, and (6) host computers. Since each component is large enough to be a research area by itself, only elements in components that are specifically related to mobile commerce are explained in detail. Lists of the technologies used for component construction are also discussed. Varshney, Vetter, and Kalakota (2000) offer related research on mobile commerce systems in their article.

Mobile Commerce Applications

The applications of electronic commerce are already widespread. For example, some tasks that are not feasible for electronic commerce, such as mobile inventory tracking and dispatching, are possible for mobile commerce. Table 1 lists some of the major mobile commerce appli-

Table 1. Major mobile commerce applications

Mobile Category	Major Applications	Clients
Commerce	Mobile transactions and payments	Businesses
Education	Mobile classrooms and labs	Schools and training centers
Enterprise resource planning	Resource management	All
Entertainment	Games/images/music/video downloads, and online gaming	Entertainment industry
Healthcare	Accessing and updating patient records	Hospitals and nursing homes
Inventory tracking and dispatching	Product tracking and dispatching	Delivery services and transportation
Traffic	Global positioning, directions, and traffic advisories	Transportation and auto industries
Travel and ticketing	Travel management	Travel industry and ticket sales

Table 2. Comparisons of two major kinds of mobile middleware

	WAP	i-mode
Developer	WAP Forum	NTT DoCoMo
Function	A protocol	A complete mobile Internet service
Host Language	WML (Wireless Markup Language)	CHTML (Compact HTML)
Major Technology	WAP Gateway	TCP/IP modifications
Key Features	Widely adopted and flexible	Highest number of users and easy to use

cations (Gordon & Gebauer, 2001; Sadeh, 2002), along with details of each.

Mobile Middleware

The term middleware refers to the software layer between the operating system and the distributed applications that interact via the networks. The primary mission of a middleware layer is to hide the underlying networked environment's complexity by insulating applications from explicit protocol handling disjoint memories, data replication, network faults, and parallelism (Geihs, 2001). Mobile middleware translates requests from mobile stations to a host computer and adapts content from the host to the mobile station (Saha, Jamtgaard, & Villasenor, 2001).

WAP and i-mode

According to an article at Eurotechnology.com (2000), 60% of the world's wireless Internet users use i-mode, 39% use WAP, and 1% use Palm middleware. Table 3 compares i-mode and WAP, along with details of each.

WAP (Wireless Application Protocol)

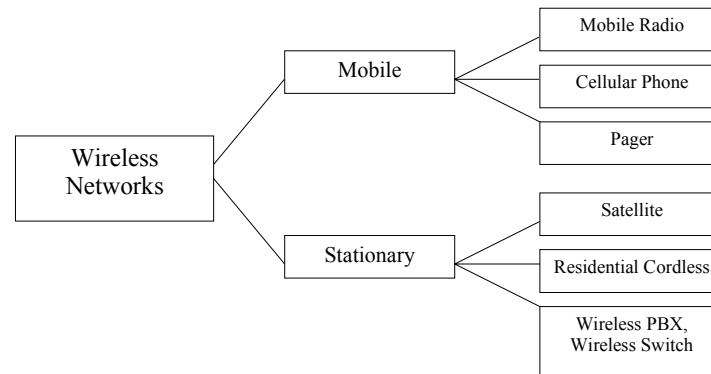
WAP (2003) is an open, global specification that allows users with mobile stations to easily access and interact with information and services instantly. It is a very flexible standard including most wireless networks, which com-

prise CDPD, CDMA, GSM, PDC, PHS, TDMA, FLEX, ReFLEX, iDEN, TETRA, DECT, DataTAC, Mobitex, and GRPS. It is supported by most operating systems and was specifically engineered for mobile stations include PalmOS, EPOC, Windows CE, FLEXOS, OS/9, and JavaOS. The most important technology applied by WAP is probably the WAP Gateway, which translates requests from the WAP protocol stack to the WWW stack, so they can be submitted to Web servers. For example, requests from mobile stations are sent as a URL through the network to the WAP Gateway; responses are sent from the Web server to the WAP Gateway in HTML and are then translated to WML and sent to the mobile stations. Although WAP supports HTML and XML, its host language is WML (Wireless Markup Language), which is a markup language based on XML that is intended for use in specifying content and user interfaces for handheld devices.

i-mode

i-mode (2003) is the full-color, always-on, and packet-switched Internet service for cellular phones offered by NTT DoCoMo. Introduced in February 1999, it has attracted over 36 million subscribers worldwide. With i-mode, cellular phone users can easily access more than 62,000 Internet sites, as well as specialized services such as e-mail, online shopping and banking, ticket reservations, and personalized

Figure 4. Categories and products of wireless communication



ringing melodies that can be downloaded for their phones. The i-mode network structure not only provides access to i-mode and i-mode-compatible contents through the Internet, but also provides access through a dedicated leased-line circuit for added security. i-mode is the only network in the world that now allows subscribers continuous access to the Internet via cellular phones. Users are charged based on the volume of data transmitted, rather than the amount of time spent connected.

Wireless Networks

Wireless networks could be divided into two major categories—stationary and mobile; as shown in Figure 4, it includes the following six types: (1) satellite networks, (2) paging networks, (3) cellular networks, (4) wireless LANs, (5) personal communication networks, and (6) packet radio networks.

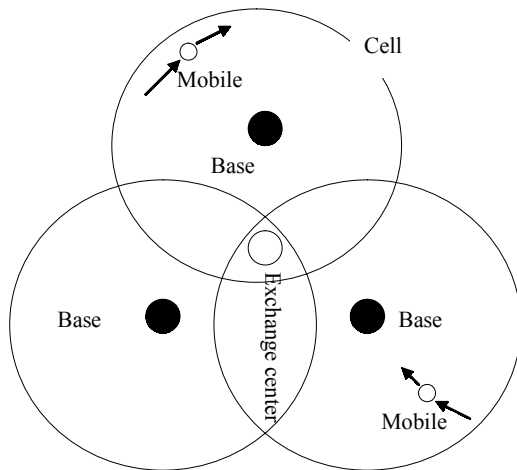
Structure of Wireless Networks

The structure of a network system is based on local area network (LAN) and wide area network (WAN). The structure of a wireless network system is similarly based on wireless local area network (WLAN), wireless wide

area network (WWAN), and wireless personal area network (WPAN).

- **Wireless Local Area Networks:** The range covered is as small as an office or a small campus. IEEE 802.11b is the protocol of WLAN. The frequency is 2.4 GHz. There is up to 11 Mbps transition rate. Besides 802.11b, IEEE 802.11a provides 5 GHz segment of frequency. It is possible to support a 54 Mbps transmitting rate.
- **Wireless Wide Area Networks:** WWAN is mainly used in cellular network systems and satellite systems. For the data transition, Cellular Digital Packet Data (CDPD) network provides 19.2 Kbps data transmitting rate.
- **Wireless Personal Area Network:** The effective range of WPAN is much smaller than both WWAN and WLAN. Normally the put though data rate is not high and the power is quite low. Using Bluetooth as example, the range covered is just the scope of a house.
- **Fixed Wireless Networks:** The fixed wireless network is applying wireless signal between communication instruments. The instruments are located at fixed posi-

Figure 5. Basic model of mobile wireless communication networking



tions. Local multipoint distribution service (LMDS) is a fixed wireless technique. It connects from one to multiple points by using frequency from 216 MHz to 600 MHz for communication. The data-transmitting rate could reach 155 Kbps. There should be no barrier on the transmitting path. The covered range of fixed wireless network is reached to a radius of 10 Km.

Figure 5 shows the basic model of mobile wireless communication networking. The ex-

change center is the vocal point to connect wired and wireless communication systems together.

Wired Networks

A wired communication could be in different functions: voice communication for telephone, and data communication for computer networks. Figure 6 shows a wired networking system.

When the needs of computer networking increase, it is possible to use a concentrator for expanding the networks with more nodes. In Figure 7, a concentrator is installed for connecting the whole computer networks. The way to further expand a computer network is by connecting multiple concentrators, as shown in Figure 8.

Host Computers

A host computer processes, produces, and stores all the information for mobile commerce applications. This component is similar to that used in an electronic commerce system because the host computers are usually not aware of differences among the targets, browsers, or micro-browsers they serve. It is the application programs that are responsible for apprehending

Figure 6. Wired networking system with one hub

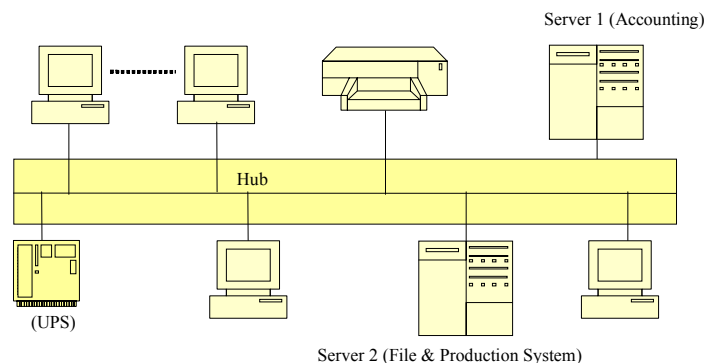


Figure 7. Wired networking system with a concentrator

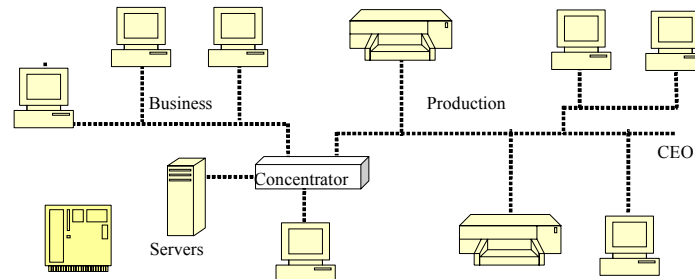
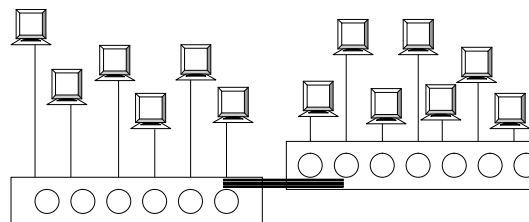


Figure 8. Expanding computer network with multiple concentrators



their clients and responding to them accordingly. Most of the mobile commerce application programs reside in this component, except for some client-side programs such as cookies. This component contains three major components: a Web server, a database server, and application programs and support software.

Web Servers

A Web server is a server-side application program that runs on a host computer and manages the Web pages stored on the Web site's database. There are many Web server software applications, including public domain software from NCSA and Apache, and commercial packages from Microsoft, Netscape, and others. Since April 1996, Apache has been the most popular HTTP server on the Internet; in May

1999, it was running on 57% of all Web servers. It was developed in early 1995 based on code and ideas found in the most popular HTTP server of the time, NCSA http 1.3. It has since evolved to rival (and probably surpass) almost any other Unix-based HTTP server in terms of functionality and speed. It features highly configurable error messages, DBM-based authentication databases, and content negotiation.

Database Servers

A database server manages database access functions, such as locating the actual record being requested or updating the data in databases. Some popular databases are Oracle9i, Microsoft Access, and IBM DB2. Other than the server-side database servers, a growing

trend is to provide a mobile database or an embedded database to a handheld device with a wide range of data-processing functionality. The functionality is frequently very sophisticated, and the flat file system that comes with these devices may not be able to adequately handle and manipulate data. Embedded databases have very small footprints, and must be able to run without the services of a database administrator and accommodate the low-bandwidth constraints of a wireless handheld network. Some leading embedded databases are Progress Software databases, Sybase's Anywhere products, and Ardent Software's DataStage (Ortiz, 2000).

Application Programs and Support Software

Web and database servers are mandatory for mobile commerce systems; application programs handle all server-side processing. However, to facilitate mobile commerce applications, some other support software is needed. For example, various programming languages, including Perl, Java, Visual Basic, and C/C++, and the CGI (Common Gateway Interface) are

necessary to transfer information between Web interfaces, and CGI scripts are necessary.

INTERNET-ENABLED MOBILE HANDHELD DEVICES

A typical system structure for handheld devices includes the following six major components: (1) a mobile operating system, (2) a mobile central processing unit, (3) a micro-browser, (4) input/output devices, (5) a memory, and (6) batteries, which will be detailed in the following sub-sections.

Mobile Operating Systems

Simply adapting desktop operating systems for mobile handheld devices has proved to be a futile endeavor; an example of this effort is Microsoft Windows CE. A mobile operating system needs a new architecture and different features in order to provide adequate services for handheld devices. Several mobile operating systems are already available, and each employs a different architecture and implementation. Figure 9 shows a generalized mobile operating system structure, which can be visualized

Figure 9. A generalized mobile operating system structure

1.	Applications
2.	Graphical user interface (GUI)
3.	Application programming interface (API) framework
4.	Multimedia Communication infrastructure Security
5.	Computer kernel Power management Real-time kernel
6.	Hardware controller

as a six-layer stack: (1) mobile applications; (2) graphical user interface (GUI); (3) application programming interface (API) framework; (4) multimedia, communication infrastructure, and security; (5) computer kernel, power management, and real-time kernel; and (6) hardware controller.

Although a wide range of mobile handheld devices are available in the market, the operating systems—the hub of the devices—are dominated by just three major organizations. The following two lists show the operating systems used in the top three brands of smart cellular phones and PDAs in descending order of market share:

- **Smart Cellular Phones:** Microsoft Smartphone 2002, Palm OS 5, and Symbian OS 7 (Vaughan-Nichols, 2003).
- **PDAs:** Palm OS 5, Microsoft Pocket PC 2002, and Symbian OS 7 (PCTechGuide, 2003).

The market share is changing frequently, and claims concerning the share vary enormously. It is almost impossible to predict which will be the ultimate winner in the battle of mobile operating systems.

Mobile Central Processing Units

The core hardware in mobile handheld devices is the mobile processors, and the performance and functionality of the devices are largely dependent on the capabilities of the processors. There used to be several brands available, but recently mobile processors designed by ARM Ltd. have begun to dominate the market. Handheld devices are becoming more sophisticated and efficient every day, and mobile users are demanding more functionality from the devices. For example, in 2002 In-Stat/MDR

predicted that worldwide mobile Internet access device unit shipments would increase from approximately 430 million that year to approximately 760 million in 2006 (In-Stat/MDR, 2002). To achieve this advanced functionality, in addition to the obvious feature, low cost, today's mobile processors must have the following features: high performance, low power consumption, multimedia capability, and real-time capability.

Micro-Browsers

Micro-browsers are a miniaturized version of desktop browsers such as Netscape Navigator and Microsoft Internet Explorer. They provide graphical user interfaces that enable mobile users to interact with mobile commerce applications. Due to the limited resources of handheld devices, micro-browsers differ from traditional desktop browsers in the following ways:

- smaller windows,
- smaller footprints, and
- fewer functions and multimedia features.

Several micro-browsers, such as Microsoft Mobile Explorer and Wapaka Java Micro-Browser, are already available. America Online, Inc., the parent company of the Netscape Network, and Nokia are developing and marketing a Netscape-branded version of Nokia's WAP micro-browser, with AOL-enhanced features, for use across a wide variety of mobile handheld devices. Figure 10 shows a typical micro-browser, a Mobile Browser version 7.0 from Openwave Systems, which includes the following features: compatibility with WAP or i-mode; multimedia support; color images and animation; and dual network stack, HTTP and WSP, support (Openwave Systems, n.d.).

Figure 10. Openwave® Mobile Browser version 7



Input/Output Devices

Mobile handheld devices have adopted various I/O devices. The only major output device is the screen, whereas there are several popular input devices, including:

- **Keyboards:** There are two kinds of keyboards: built-in keyboards and external, plug-in keyboards. The problem with the former is that they are too small for touch-typing, whereas the latter suffers from inconvenience. Fabric keyboards that can be rolled up or folded around the handheld devices are being developed to relieve the problem of external keyboards.
- **Touch Screens/Writing Areas with a Stylus:** A touch screen is a display that is sensitive to human touch, allowing a user to interact with the applications by touching pictures or words on the screen, and a stylus is an input device used to write text or draw lines on a surface as input to a handheld device.

Memory

Desktop PCs or notebooks usually have between 64 to 256 Mbytes of memory available

for users, whereas handheld devices typically have only 4 to 64 Mbytes. PDAs normally have more storage space than smart cellular phones. The former commonly have 16 Mbytes, and the latter may have a memory size as low as a few Kbytes. Handheld devices usually employ three types of memory: (1) random access memory (RAM), (2) read-only memory (ROM), and (3) flash memory. It is expected that handheld devices will adopt hard disks, which provide much more storage capacity, in the near future. A comprehensive survey of storage options can be found in Scheible (2002).

Batteries

Rechargeable Lithium Ion batteries are the batteries most commonly used by handheld devices. The life of this kind of battery is short, generally only a few hours of operating time. Battery technology will not significantly improve unless and until manufacturers begin to switch to fuel cells, which is unlikely in the near future. A fuel cell operates like a battery, but unlike a battery, a fuel cell does not run down or require recharging and will continue to produce energy in the form of electricity and heat as long as fuel is supplied. Since the fuel cell relies on chemical energy rather than combustion, emissions would be much lower than emissions from the cleanest existing fuel combustion processes.

Synchronization

Synchronization connects handheld devices to desktop computers, notebooks, and peripherals in order to transfer or synchronize data. The traditional method of synchronization uses serial cables to connect handheld devices and other computing equipment. Now, however, many handheld devices use either an infrared (IR) port or Bluetooth technology to send infor-

mation to other devices without needing to use cables:

- IrDA Data (Infrared Data Association, n.d.), a standard formulated by the Infrared Data Association to ensure the quality and interoperability of infrared hardware, is designed for data transfer over distances of up to one meter, acting as a point-to-point cable replacement.
- Bluetooth wireless technology is a specification aiming at simplifying communications among handheld devices, printers, computers, and other devices based on short-range radio technology. The Bluetooth 1.1 specifications (Bluetooth SIG, n.d.) consist of two documents: the Core, which provides design specifications, and the Profile, which provides interoperability guidelines.

MOBILE COMMERCE PAYMENT METHODS

Regardless of the bright future of mobile commerce, its prosperity and popularity will be brought to a higher level only if information can be securely and safely exchanged among end systems (mobile users and content providers). Applying the security and payment technologies for electronic commerce to mobile commerce has been proven to be a futile effort because electronic commerce and mobile commerce are based on different infrastructures (wired versus wireless). A wide variety of security procedures and payment methods, therefore, have been developed and applied to mobile commerce. These technologies are extremely diverse and complicated, and a satisfactory discussion on them is still absent. This section attempts to provide an introduction of mobile commerce payment methods.

REQUIREMENTS AND PROPERTIES OF M-COMMERCE SYSTEMS, SECURITY, AND PAYMENT METHODS

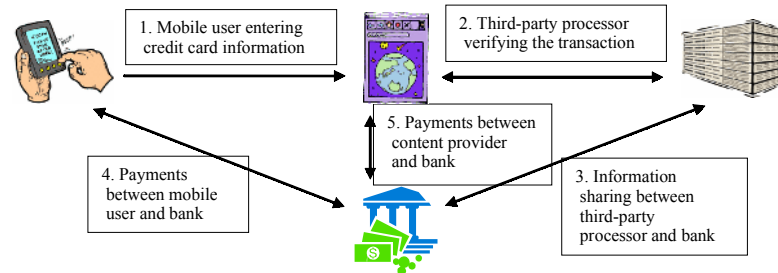
It is first necessary to examine what kind of features mobile commerce security is expected to have in order to conduct effective and efficient mobile commerce transactions, and what kind of challenges may be faced in the process of developing new mobile commerce security. The requirements and properties of secure mobile commerce information and systems are as follows (Lee, Kou, & Hu, 2004):

- **Confidentiality:** The information and systems must not be disclosed to unauthorized persons, processes, or devices.
- **Authentication:** Ensures parties to a transaction are not impostors and are trusted.
- **Integrity:** The information and systems have not been altered or corrupted by outside parties.
- **Authorization:** Procedures must be provided to verify that the user can make the requested purchases.
- **Availability:** An authorized user must have timely, reliable access to information in order to perform mobile commerce transactions.
- **Non-Repudiation:** Ensures a user cannot deny that he or she performed a transaction; the user is provided with proof of the transaction, and the recipient is assured of the user's identity.

The requirements for mobile commerce security are:

- Confidentiality, authentication, integrity, authorization, availability, and non-repudiation must be rigorously enforced.

Figure 11. A typical macro-payment scenario



- They should be interoperable for most systems.
- They should be acceptable by the current or future systems with reduced cost.
- No mobile commerce transactions are deferred or deterred because of the deployment.
The requirements for mobile payment methods are the same as the ones of mobile commerce security with an additional item:
- They should allow content providers to provide affordable, easy-to-use, efficient, and interoperable payment methods to users.

Macro-Payment and Micro-Payment Methods

Mobile commerce payment methods are the means used to pay for goods or services with a mobile handheld device, such as a smart cellular phone or an Internet-enabled PDA. There are usually two kinds of mobile commerce payment methods:

- **Macro-Payments:** Used by traditional electronic commerce and usually involve amounts more than US\$10. Payments by

credit cards are the most common method for macro-payments.

- **Micro-Payments:** Usually involve amounts less than US\$10, which are too small to be economically processed by credit cards. The amounts are usually charged to users' phone bills.

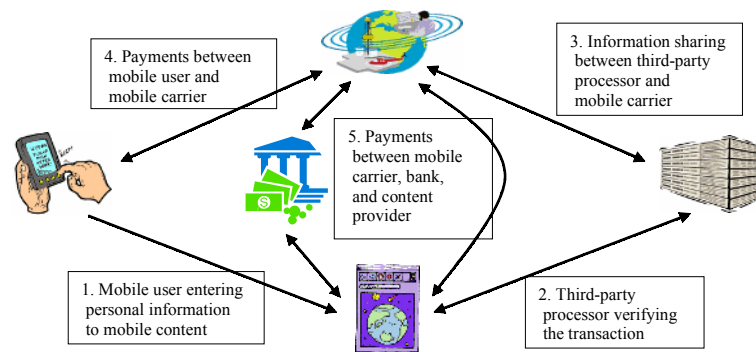
A typical macro-payment scenario is as follows and is illustrated in Figure 11:

1. A mobile user submits his/her credit card information to the mobile content via a handheld device.
2. A third-party processor verifies and authorizes the transaction.
3. The third-party processor routes verification and authorization requests to the card issuing bank.
4. The user pays his/her monthly credit card bill.
5. The bank pays the mobile content provider for the user.

On the other hand, a typical micro-payment scenario is as follows and is illustrated in Figure 12:

1. A mobile user submits his/her personal information to the mobile content via a handheld device.

Figure 12. A typical micro-payment scenario



2. A third-party processor verifies and authorizes the transaction.
 3. The third-party processor routes verification and authorization requests to the mobile carrier.
 4. The user pays his/her monthly phone bill.
 5. The mobile carrier pays the mobile content provider directly or through a bank after deducting transaction fees for the user.
1. **Mobile Commerce Applications:** Electronic commerce applications are already broad. Mobile commerce applications not only cover the existing applications, but also include new applications, which can be performed at any time and from anywhere by using mobile computing technology.
 2. **Internet-Enabled Mobile Handheld Devices:** Handheld devices are limited by their small screens, limited memory, limited processing power, and low battery power, and suffer from wireless network transmission problems. Numerous handheld devices, such as PDAs or Web-enabled cellular phones, are available in the market, but most use one of three major operating systems: Palm OS, Microsoft Pocket PC, and Symbian OS. At this moment, Palm OS leads the market, although it faces a serious challenge from Pocket PC.

CONCLUSION AND FUTURE DIRECTIONS

This section gives conclusions and future directions of the three mobile commerce themes discussed in this chapter: mobile commerce systems, Internet-enabled mobile handheld devices, and mobile payment methods.

Mobile Commerce Systems

A mobile commerce system involves a range of disciplines and technologies. This level of complexity makes understanding and constructing a mobile commerce system an arduous task. To facilitate this process, this chapter divided a mobile commerce system into six components, which can be summarized as follows:

3. **Mobile Middleware:** WAP and i-mode are the two major kinds of mobile middleware. WAP is widely adopted and flexible, while i-mode has the highest number of users and is easy to use. It is difficult to predict which middleware will be the eventual winner in the end; it is

more likely that the two will be blended somehow at some point in the future.

4. **Wireless and Wired Networks:** Wireless communication capability supports mobility for end users in mobile commerce systems. Wireless LAN, MAN, and WAN are major components used to provide radio communication channels so that mobile service is possible. In the WLAN category, the Wi-Fi standard with 11 Mbps throughput dominates the current market. It is expected that standards with much higher transmission speeds, such as IEEE 802.11a and 802.11g, will replace Wi-Fi in the near future. Compared to WLANs, cellular systems can provide longer transmission distances and greater radio coverage, but suffer from the drawback of much lower bandwidth (less than 1 Mbps). In the latest trend for cellular systems, 3G standards supporting wireless multimedia and high-bandwidth services are beginning to be deployed. WCDMA and CDMA2000 are likely to dominate the market in the future.
5. **Host Computers:** Host computers process and store all the information needed for mobile commerce applications, and most application programs can be found here. They include three major components: Web servers, database servers, and application programs and support software.

Internet-Enabled Mobile Handheld Devices

Internet-enabled mobile handheld devices are one of the core components of mobile commerce systems, as they are needed for mobile users to directly interact with mobile commerce applications. Understanding the devices and knowing their functions and capabilities is vital

for the success of mobile commerce applications. A handheld device relies on a wide range of disciplines and technologies for its success. To facilitate the understanding, this chapter broke down the functions of a handheld device into six major components, which can be summarized as follows:

1. **Mobile Operating Systems:** Simply adapting desktop operating systems for handheld devices has proved to be futile. A mobile operating system needs a completely new architecture and different features to provide adequate services for handheld devices. A generalized mobile operating system structure can be visualized as a six-layer stack: (1) applications; (2) GUI; (3) API framework; (4) multimedia, communication infrastructure, and security; (5) computer kernel, power management, and real-time kernel; and (6) hardware controller.
2. **Mobile Central Processing Units:** Handheld devices are becoming more sophisticated and efficient everyday, and mobile users are demanding more functionality from their devices. To achieve this advanced functionality, in addition to the obvious feature, low cost, today's mobile processors must have the following features: high performance, low power consumption, multimedia capability, and real-time capability. The cores and architectures designed by Cambridge-based ARM Holdings Ltd. have begun to dominate the mobile CPU market.
3. **Micro-Browsers:** Micro-browsers are miniaturized versions of desktop browsers such as Netscape Navigator and Microsoft Internet Explorer. They provide graphical user interfaces that allow mobile users to interact with mobile commerce applications. Micro-browsers usu-

ally use one of the following approaches to return results to the mobile user: wireless language direct access, HTML direct access, HTML to wireless language conversion, and error.

4. **Input/Output Devices:** Mobile handheld devices have adopted various I/O devices. The only major output device is the screen, but there are several popular input devices, which are keyboards and touchscreens/writing areas that need a stylus.
5. **Memory:** Three types of memory are usually employed by handheld devices: RAM, ROM, and flash memory. Hard disks, which provide much more storage capacity, are likely to be adopted by handheld devices in the near future.
6. **Batteries:** At present, rechargeable Lithium Ion batteries are the most common batteries used by handheld devices.

However, the life of this kind of battery is short and the technology will not significantly improve unless and until manufacturers begin to switch to fuel cells, which may not happen for at least several years.

Synchronization connects handheld devices to desktop computers, notebooks, or peripherals to transfer or synchronize data. Without needing serial cables, many handheld devices now use either an infrared (IR) port or Bluetooth technology to send information to other devices.

Mobile handheld devices are usually divided into two types: smart cellular phones and Internet-enabled PDAs (personal digital assistants). These two kinds of devices started out as very different products, but they have gradually blended into each other. In the near future, it will be difficult to tell the difference between these two types of devices. The newest prod-

Table 3. Specifications of some major mobile handheld devices

Vendor & Device	Operating System	Processor	Installed RAM/ROM	Input Methods	Key Features
Compaq iPAQ H3870	MS Pocket PC 2002	206 MHz Intel StrongARM 32-bit RISC	64 MB/32 MB	Touchscreen	Wireless e-mail/Internet
Handspring Treo 300	Palm OS 3.5.2H	33 MHz Motorola Dragonball VZ	16 MB/8 MB	Keyboard/stylus	CDMA network
Motorola Accompli 009	Wisdom OS 5.0	33 MHz Motorola Dragonball VZ	8 MB/4 MB	Keyboard	GPRS network
Nokia 9290 Communicator	Symbian OS	32-bit ARM9 RISC	16 MB/8 MB	Keyboard	WAP
Nokia 6800	Series 40			Keyboard	Innovative keyboard integration
Palm i705	Palm OS 4.1	33 MHz Motorola Dragonball VZ	8 MB/4 MB	Stylus	Wireless e-mail/Internet
Samsung SPH-i330	Palm OS 4.1	66MHz Motorola Dragonball Super VZ	16 MB/8 MB	Touchscreen/Stylus	Color screen
Sony Clie PEG-NR70V	Palm OS 4.1	66 MHz Motorola Dragonball Super VZ	16 MB/8 MB	Keyboard/stylus/touchscreen	Multimedia
Sony Ericsson T68i			800KB	Keyboard	Multimedia messaging service
Toshiba E740	MS Pocket PC 2002	400 MHz Intel PXA250	64 MB/32 MB	Stylus/Touchscreen	Wireless Internet

ucts such as tablet PCs belong to the category of PDAs because both have similar functionality. There are numerous mobile devices available in the market today. Table 3 lists some major mobile device specifications, although several table entries are incomplete, as some of the information is classified as confidential due to business considerations.

Mobile Payment Methods

Another important issue for mobile commerce is mobile security and payment. Mobile commerce systems can prosper only if information can be securely exchanged among end systems (consumers and vendors). Security issues (including payment) include data reliability, integrity, confidentiality, and authentication, and are usually an important part of implementation in wireless protocols/systems. Solutions are updated frequently, due to the lack of a comprehensive wireless security infrastructure and standard. A unified approach has not yet emerged. Among the many themes of mobile commerce security, mobile payment methods are probably the most important. These consist of the methods used to pay for goods or services with a mobile handheld device, such as a smart cellular phone or an Internet-enabled PDA. A typical mobile payment process includes: registration, payment submission, authentication and authorization by a content provider, and confirmation.

REFERENCES

Bluetooth SIG. (n.d.). *Bluetooth specifications*. Retrieved August 12, 2004, from <https://www.bluetooth.org/foundry/specification/document/specification>

Cormie, D. (2002). *The ARM11 microarchitecture*. Retrieved July 21, 2004,

from [http://www.arm.com/support/59XGYS/\\$File/ARM11+Microarchitecture+White+Paper.pdf](http://www.arm.com/support/59XGYS/$File/ARM11+Microarchitecture+White+Paper.pdf)

Eurotechnology.com. (2000). *Frequently asked questions about NTT-DoCoMo's i-mode*. Retrieved December 16, 2002, from <http://www.eurotechnology.com/imode/faq.html>

Geihs, K. (2001). Middleware challenges ahead. *IEEE Computer*, 34(6), 24-31.

Gordon, P., & Gebauer, J. (2001). *M-commerce: Revolution + inertia = evolution*. Working Paper 01-WP-1038, University of California—Berkeley, California.

Hu, W., Lee, C., & Yeh, J. (2003). Mobile commerce systems. In N. Shi (Ed.), *Mobile commerce applications* (pp. 1-23). Hershey, PA: Idea Group Inc.

i-mode. (2003). *NTT-DoCoMo*. Retrieved November 28, 2002, from <http://www.nttdocomo.com/>

Infrared Data Association. (n.d.). *Technical summary of "IrDA DATA" and "IrDA CONTROL"*. Retrieved July 15, 2004, from <http://www.irda.org/standards/standards.asp>

In-Stat/MDR. (2002). *Demand increasing for mobile Internet access devices—handsets represent primary growth driver*. Retrieved July 8, 2004, from <http://www.instat.com/press.asp?ID=250&sku=IN020280MD>

Leavitt, N. (2000). Will WAP deliver the wireless Internet? *IEEE Computer*, 34(5), 16-20.

Leavitt, N. (2003). Will wireless gaming be a winner? *IEEE Computer*, 36(1), 24-27.

Microsoft Corporation. (2003a). *Pocket PC*. Retrieved June 25, 2004, from <http://www.microsoft.com/windowsmobile/products/pocketpc/default.mspx>

- Microsoft Corporation. (2003b). *Smartphone*. Retrieved June 23, 2004, from <http://www.microsoft.com/windowsmobile/products/smartphone/default.mspx>
- PCTechGuide. (2002). *Mobile computing*. Retrieved November 2, 2002, from <http://www.pctechguide.com/25mobile.htm>
- Openwave Systems. (n.d.). *Mobile browser V7*. Retrieved July 15, 2004, from http://www.openwave.com/products/device_products/phone_tools/mobile_browser_7.html
- Ortiz, S. Jr. (2000). Embedded databases come out of hiding. *IEEE Computer*, 33(3), 16-19.
- Pahlavan, K., & Krishnamurthy, P. (2002). *Principles of wireless networks: A unified approach*. Upper Saddle River, NJ: Prentice-Hall.
- Palm Source. (n.d.). *Palm OS*. Retrieved December 22, 2004, from <http://www.palmsource.com/palmos/>
- PCTechGuide. (n.d.). *Mobile computing*. Retrieved July 2, 2004, from <http://www.pctechguide.com/25mobile.htm>
- Reuters. (2001). *The Yankee Group publishes U.S. mobile commerce forecast*. Retrieved June 16, 2004, from http://about.reuters.com/newsreleases/art_31-10-2001_id765.asp
- Sadeh, N. (2002). *M-commerce: Technologies, services, and business models* (pp. 177-179). New York: John Wiley & Sons.
- Saha, S., Jamtgaard, M., & Villasenor, J. (2001). Bringing the wireless Internet to mobile devices. *IEEE Computer*, 34(6), 54-58.
- Scheible, J. P. (2002). A survey of storage options. *IEEE Computer*, 35(12), 42-46.
- Symbian. (2003). Retrieved December 10, 2002, from <http://www.symbian.com/>
- Varshney, U., Vetter, R. J., & Kalakota, R. (2000). Mobile commerce: A new frontier. *IEEE Computer*, 33(10), 32-38.
- Vaughan-Nichols, S. J. (2003). OSs battle in the smart-phone market. *IEEE Computer*, 36(6), 10-12.
- WAP (Wireless Application Protocol). (2003). *Open Mobile Alliance Ltd*. Retrieved November 21, 2002, from <http://www.wapforum.org/>
- The Yankee Group. (2001). *Over 50% of large U.S. enterprises plan to implement a wireless/mobile solution by 2003*. Retrieved December 10, 2002, from http://www.yankee-group.com/public/news_releases/news_release_detail.jsp?ID=PressReleases/news_09102002_wmec.htm

Section VI

Strategy

Chapter XXIX

Strategic Perspectives in Mobile Banking: Technology, Value Creation, and Developing Factors

Achraf Ayadi

GET/Institut National des Télécommunications, France

Chantal Ammi

GET/Institut National des Télécommunications, France

ABSTRACT

The convergence of the Internet and mobile networks has created new opportunities and applications. Considering mobile business only as an extension of the traditional web can lead to missing out on unique and differentiable qualities for new value-added opportunities. Mobile banking is considered as potentially one of the most value-added and important mobile service available. The chapter examines the technological changes in mobile networks and the innovative attributes of mobile Internet. It advances the theoretical framework of innovation in services to develop a customer centric analysis of m-banking value proposition. The chapter goes on to discuss critical factors in the diffusion of m-Banking and explores reasons of failure and further prospects of success.

INTRODUCTION

The mobile communications market is changing, and the next generation of customers will require more than vocal services. The technological and commercial convergence of mobile networks and Internet finds telecom operators faced with new challenges and enormous opportunities (Booz Allen & Hamilton, 2000).

Thus, in spite of the growth of user numbers and the growing traffic on mobile networks, the rise of competition has led to a strong fall in prices and margins. Differentiating products must be realized less for vocal communications than for data exchanges (IDATE, 2003). The traditional income of telecom operators—initially based on relatively constant subscription fees—will yield more place to economic models based on

mobile Internet. Then, new incomes could emerge from subscriptions to services like data and contents, m-commerce, advertising, and advanced network services like virtual private networks (VPNs) and quality of service (QoS) guarantees. This modification of competition basis in the mobile market is accentuated by deep changes in consumer behavior: the Internet has caused a quick evolution of needs moving from social communication to electronic commerce. So, the convergence of mobile communications and the Internet requires a new analysis of the current model of value creation.

Mobile banking (m-banking) is considered as one of the most important emerging services implying actors from different economic sectors in the m-commerce value chain (Lee, McGoldrick, Keeling, & Doherty, 2003; Celent Communications, 2002; Barnes, 2002). “M-banking” consists of managing a bank account through a wireless Internet-enabled device.

On the basis of the innovation diffusion theory in service sectors, we study the technological tendencies in mobile Internet. Then, we will analyze the development factors of mobile services in the banking industry and their impact on the value chain. Finally, we will conclude with strategic perspectives of mobile banking and its future evolution.

TECHNOLOGICAL CHANGES IN MOBILE NETWORKS

Trends in Mobile Network Evolution: An Overview

Mobile technologies have known many standards:

- The first generation of mobile networks (1G) was based on voice exchange via analog radio frequencies.
- The second generation (2G) is fragmented between IS-54 and IS-95 US’ standards and GSM (Global System for Mobile), which is the most expanded standard (50% of the market share in the world (IDATE, 2003)). However, the 2G standards in general suffer from low capacities in data transmission. Thus, they have been quickly supplanted by standards known as 2.5G (HSCSD, GPRS, and EDGE) which improve data transfer significantly. This step was accompanied by two forms of mobile Internet: WAP and i-mode. The WAP (wireless application protocol) is a protocol without license which has been very quickly adopted by telecom operators and the equipment industry of mobile phones as an access bridge to Web pages throughout a mobile telephone. i-mode was launched in February 1999 by NTT DoCoMo with more than 36.7 million subscribers in Japan (Benchmark Group, 2002). Its popularity is due to the similarity between development language i-mode, the C-HTML (Compact Hypertext Markup Language), and the Internet’s HTML. This resemblance makes possible for subscribers an easy access to Internet services, and facilitates the use of e-mail and the access to thousands of Web pages. Today, NTT DoCoMo—with participation by several mobile operators in Germany, the United Kingdom, and France—is trying to settle into Europe where i-mode is emerging as a benchmarking standard. This intermediate stage of 2.5G standards has allowed the operators and various other actors in content creation to come near a dubious prospects market.
- The heavy investment in third-generation (3G) licenses for the Universal Mobile Telecommunication System (UMTS) standard is a brake for mobile Internet growth.

UMTS networks boost data exchange flow to comparable levels known in traditional PC Internet connections. The services considered as pioneers are expensive and need to focus on narrow niche customers (innovators, professional customers, etc.). For example, the French mobile operator “SFR” offers only one tariff formula: 349 euros to acquire a wireless PC access card and a subscription of 75 euros per month for 10 hours of connection with a limit of 500 Mo of data exchange volume. Admittedly, the tests carried out seem satisfactory on the connection quality, but such offers are limited by a fundamental aspect: the difficulty to convince the broad market of millions of individual users of 3G services in a context when the compatible apparatuses with UMTS standard are rare and expensive.

In addition to these changes in mobile networks, handheld devices (principally handhelds and mobile phones) are converging to combining communication and data management functionalities. Memory size improvements and processing capabilities of mobile devices are boosting these changes.

The Future of Mobile Internet: Technology, User Communities, and Applications

Uncertainties are numerous on the future of mobile networks (van der Kar & van der Duin, 2004). However, the projections of the number of potential subscribers remain very favorable, with an estimation around 320 million subscribers in the world by 2006 (Benchmark Group, 2003a; IDATE, 2003). A study undertaken by Lehmann, Kuhn, and Lehner (2004) released the most significant tendencies on the levels of technologies, with user communities and appli-

cations becoming essential in the future as m-commerce keystones. Unless there are many regularities in their findings, Lehmann et al. (2004) agree with van der Kar and van der Duin (2004) about the fact that the opinions of questioned experts at the time of their respective investigations remain divergent on various aspects. Although the mobile park is increasing and the rate of penetration of mobile equipment is higher than 65%, the demand structure is not really developed and the number of small “packages” lower than 2h or prepaid charts represents the majority of the users with few perspectives of evolution. The commercialization of high-priced “packages,” including advanced Internet mobile services, seems difficult. Most of the mobiles operators are in the following configuration: as multi-operators (fixes, mobiles, Internet), they undergo a demand cannibalization (between their alternative networks) and insufficient resources because of the nature of deployed networks and the underdeveloped and saturated demand in mobile network. More than that, the mobiles operators expect in the long term a cannibalization between the two types of networks (2.5 and 3) and are faced with an insoluble question: would the current customers of GSM, GPRS, or i-mode will become UMTS customers in the future? If yes, in which form and in which proportions? The only certainty in this matter is the necessity to maintain two incompatible networks for the operators and the incapacity for a same consumer to use two different networks at the same time.

Uncertainties on the future and technological constraints of the current networks (2G or 2.5G) make hypothetical the development of the services of mobile Internet. Access providers and service suppliers, conditioned by these facts, seem very careful and wait for “real” opportunities not already known.

This technological context in transition affects the development of trade via the mobile networks, delays the access to rich multimedia contents, and accentuates the questions around “the value proposition” to offer via mobile Internet.

INNOVATION AND VALUE CREATION IN MOBILE INTERNET SERVICES

Innovation in Services

In the literature, we identify three main processes to carry out innovation projects: the decision process (innovation strategies), the executing process (organizing innovation), and the diffusion process (confronting innovation and market) (Loilier & Tellier, 1999). Rogers’s innovation theory presents diffusion as a process of progressive communication between the members of a social system (Rogers, 1983). Referring to an epidemiologic model, the propagation of innovation depends of the number of adopters and the importance of their communication with the rest of the population. The perception of costs and advantages of the innovations by potential adopters determines this process of diffusion (Loilier & Tellier, 1999). Thus, Rogers shows that this perception is closely related to intrinsic characteristics of innovation on the basis of which the product will

be appreciated. Some authors define the innovation as the existence of “creative attributes of value” (Filipo, 2000). In service innovation literature, the innovation is considered as a customer-oriented value creation process. Then, concepts like “innovation” and “value proposition” are dependent, interrelated, and inseparable to analyze and determine success factors in introducing a new product (Filipo, 2000).

Value Creation in Mobile Internet

“Value propositions define the relationship between the supplier offerings and the consumer purchases by identifying how the supplier fulfills the customer needs across the different customer roles” (Clarke, 2001). Thus, a purchasing behavior via mobile Internet is different from the one via PC because the customer refuses a long navigation to search for a product or a service. In m-commerce, the customer is waiting for a personalized offer, targeted on products in immediate vicinity and accessible services beyond time and space. Some estimation evaluates that every additional click on a mobile terminal reduces the probability of the transaction by 50% (Clarke, 2001). Four value propositions in m-commerce applications are identified: ubiquity, convenience, localization, and personalization (see Table 1).

Many analyses have considered the diffusion process as the most important determinant of success in the adoption of mobile Internet

Table 1. Value proposition and applications in m-commerce (Clarke, 2001)

Suggested Value		Applications
Ubiquity	Available products everywhere, every time	News, sports scores, stock prices, travel, and weather info
Convenience	Eliminating PC connection constraints	Entertainment, gaming, retailing, banking, payments
Localization	Offer product and services depending on geographic localization	Dispatch, scheduling, discounting, emergency services, supply chain management
Personalization	Mobile phones are individual devices enabling one-to-one marketing	Advertising, database development, knowledge management systems

(Booz Allen & Hamilton, 2000; Celent Communications, 2002; IDATE, 2003). However, uncertainties on the prospective growth of mobile services in general, customer's response, and time in particular needed by 3G networks to replace 2G and 2.5G networks are very important (van der Kar & van der Duin, 2004). These uncertainties are strongly related to technologies' lifecycle replacement and breaking technologies (Loilier & Tellier, 1999). In fact, a performing technology can fail to replace another one because of a long time-to-market process or an existence of a largely diffused and established old technology (Shapiro & Varian, 1999). This failure can be related to cultural facts. Large differences are observed in the penetration rates of mobile Internet in the world (72.3% in Japan, 59.1% in Korea vs. 16.5% in Finland and 5.6% in France (Benchmark Group, 2003b)). The success of i-mode in Japan, for example, is related to particular socio-cultural factors (a better acceptability of the technology by the population) which affect the interpersonal communication (Ishii, 2004), the specificities of the Japanese networks (number of mobile greater than the number of fixed networks, size of the network smaller according the superficies of the country, and a higher demographic density).

Thus, with the growing implication of non-telecom actors in the innovation process, the mobile Internet value chain is changing dramatically (Barnes, 2002; Buellingen & Woerter, 2004). In this context, the analysis of value proposition needs to integrate multiple value systems and different value chains. A recent study of Van de Kar and Van der Duin (2004) has shown that the only certainties around the future of mobile Internet are three: the mobile payments will profit to a large accessibility, mobile services will be "multimedia," and they will privilege the data exchanges as much as voice communication.

DEVELOPING M-BANKING: A CUSTOMER-CENTRIC VALUE

Confronting Customer Needs and M-Banking Applications

Pousttchi and Schurig (2004) identify four use cases of m-banking: request of account balance, control of account movements, instant payment, and account administration. These usage cases correspond to specific customer needs and depend more on the wireless device availability than the telecommunication network performance. We can distinguish three principal applications for m-banking.

SMS-Banking

Short Message Service (SMS) is particularly adapted to 2G networks because they require low capabilities for data exchange (160 characters for 7 bits by message). For example, SMS makes possible the quick answer to a customer request to the consult of his balance account. The sensitivity of this information requires that the bank lodges and manages its own SMS server since the telecom operators are not authorized to treat them. In the United Kingdom, First Direct attracted 138,000 subscribers with its "SMS alerts" which represented 25% of its online customers (Sangani, 2002). "The main problem with this kind of transmission is the missing encryption of the data during the on-the-air transmission between the service center and the mobile phone" (Pousttchi & Schurig, 2004). Thus, banks are just satisfied to deliver limited information service to customers, but it is impossible to carry out SMS-based transactions.

WAP-Banking

Wireless application protocol (WAP) offers access to micro-Web sites managed by a bank's

Table 2. Corresponding offered services to m-banking applications

Banking Applications	Potential Offered Services of M-Banking				
		Request of account balance	Control of account movements	Instant m-payments	Account administration
	E-banking	*	**	***	***
	SMS banking	***	**	*	*
	WAP/i-mode banking	**	***	**	*
	Java BANKING	*	**	***	***

NB: (*) = low potential, (**) = medium potential, (***) = high potential

NB: (*) = low potential, (**) = medium potential, (***) = high potential

server. The customer access process resembles that of the Internet. With transactions, safety via WAP is guaranteed by Internet cryptography systems. The customer authentication is made via a PIN (personal identification number) code and the transaction authorization is given by a customer validation (TAN: Number Transaction). However, WAP-banking experience is considered as a failure for multiple reasons (Sangani, 2002): WAP requires 30-40 seconds of connection login, an important number of “clicks” before accessing useful information or carrying out a full transaction.

Java-Banking

The Java programming language is completely independent from any platform. With the proviso that there is enough space in the mobile device to store Java application, a http connection can be made with SSL-encryption protocol (Pousttchi & Schurig, 2004). In this case, accounts data can be accessed directly to banking Web servers. Java-banking presents many advantages, especially off-line usage, and interoperability and compatibility with many offered services. It is probably the nearest m-banking system to e-banking (see Table 2).

Admittedly, several financial institutions like Nordea, Sampo-Leonia, SEB, Egg, and Credit Suisse use the WAP experience to test the market and to develop a broad pallet of

functionalities (Research and Markets, 2001). Waiting for a generalization of 3G networks and new technical possibilities in terms of flow capabilities and content richness, the current availabilities are too much limited, unsecured, and unstable for banks (Lee et al., 2004).

The fast development of technologies and the announcement of new powerful applications generate the prudence of banks' leaders and increase uncertainties around the future of m-banking (Pousttchi & Schurig, 2004). The key element to choosing the best time for adoption (by the banks) and the opportunity to invest in m-banking is to anticipate the value propositions awaited by customers.

Value Propositions as Developing Factors of M-Banking

In the banking industry, distant channels of distribution are increasing (ATM, Internet, mobile phones, PDAs, etc.), but they do not seem to meet the same needs. Each channel offers different value creation opportunities to customers. For example, there are very strong differences between the Internet via PC and mobile Internet. M-banking, compared with Internet banking (e-banking), requires more personalization because of a the profile of customer and the fact that his localization is more detectable. However, it is necessary to establish fast access to information because

the customer refuses a long navigation via his mobile phone (Research and Markets, 2001; Sangani, 2002). Table 3 offers a comparison of the plausible gap of value for e-banking and m-banking (Clarke, 2001).

The Internet and Internet-enabled wireless devices could function as complementary channels of distribution (Research and Markets, 2001). They probably direct to the same customer, but they do not direct to the same needs. Thus, it seems natural that a successful bank such as Barclays (in the UK) with 1.9 million e-banking customers invests in mobile channels of distribution (Sangani, 2002). The adoption factors of m-banking development are not only related to the commercial client relationship. Many other actors affect the value chain of m-commerce whose diffusion depends upon their coordination and consensus around common technological standards, as well as their commitment in a coordinated step of market preparation (MobeyForum, 2003).

CONCLUSION AND FUTURE DIRECTIONS

The opportunity of mobile phones to create new types of value is very important in banking services. However, the current penetration of m-banking remains very mitigated. An investigation of Benchmark Group with 1,187 Net surfers (95% of them use e-banking) showed that 65% do not use m-banking and 35% are not

satisfied by offered services (Benchmark Group, 2003b). Also, Wells Fargo, one of the first banks to set up an e-banking services in the world, has stopped its m-banking services after only 2,500 customers subscriptions in a year. Many other financial institutions like the Bank of Montreal and Citibank shut down their m-banking services in mid-2002. Only the basic low-cost wireless services built around text alerts and e-mail messages have been maintained because of slow adoption rates. Other factors, like a large Internet connectivity and a superior user experience of e-banking, have reduced the need for m-banking services. This relative failure is quasi unanimous in the academic literature, but the majority of the contributors underline the still unexplored potential of m-banking. In fact, uncertainties affect specific relational and technological factors to the m-commerce market (MobeyForum, 2003). Thus, many actions remain to be done for the development of electronic payment standard, which depends on a large acceptance between value chain actors. Payment security, for example, is one of the major uncertainties in m-commerce because of the keen competition carried out by the US-UE consortium "PayCircle," between Hewlett Packard, Lucent Technologies, Oracle, Sun Microsystems, and Siemens (Sangani, 2002). The harmonization of security standards is underlined by many studies as being an accelerator element of m-banking adoption and diffusion (Lee et al., 2003).

Table 3. Corresponding offered services to m-banking applications

	E-Banking	M-Banking	Superior Gap of Value for...
Speed	*****	*	E-Banking
Price	*****	**	M-Banking
Ubiquity	*	*****	
Convenience	**	*****	
Localization	**	*****	
Personalization	***	****	

NB: (*) = weakly valuable, (*****) = highly valuable

However, some interesting initiatives are developing to federate technological visions and to reorganize competition in a “win-win” global approach. Such is the case of the “MobeyForum,” a think-tank founded in May 2000 by leading financial institutions and actors of the telecom industry, and joined by technology operators and consultants (20 members). In a recent study, this group underlines the importance of several factors to promote a true market of financial services at the international level: the development of compatible and interoperable standards to keep an open market for customers and actors, the guarantees of the independence between actors and standards to let customers choose the best combination “operator-financial institutions,” and the integration between the actual technologies into the new platforms and standards under development to avoid spending more money and time—especially for banks—to install new electronic channels of distribution (MobeyForum, 2003).

The challenge still open for all these actors is to build adapted products and services with customer needs, to shorten the time-to-market of new technologies, and to guarantee that the access cost would not be perceived by the market as a barrier to entry.

REFERENCES

- Barnes, S. J. (2002, April). The mobile commerce value chain: Analysis and future developments. *International Journal of Information Management*, 22(2), 91-108.
- Benchmark Group. (2002). *Japon: Le marché de l'Internet Mobile*. Retrieved June 15, 2004, from www.journaldunet.com
- Benchmark Group. (2003a). *Monde: Marché de l'Internet Mobile*. Retrieved June 15, 2004, from www.journaldunet.com
- Benchmark Group. (2003b). *Les internautes et leur banque: Services mobiles, mention passable*. Retrieved June 15, 2004, from www.journaldunet.com
- Booz Allen & Hamilton. (2000). The wireless Internet revolution. *Insights*, 6(2).
- Buellingen, F., & Woerter, M. (2004, December). Development perspectives, firm strategies and applications in mobile commerce. *Journal of Business Research*, 57(12), 1402-1408.
- Celent Communications. (2002, June). *Open finance in the U.S. and Europe: A perfect storm in retail banking*. New York: Celent Communications.
- Clarke, I. III. (2001). Emerging value propositions for m-commerce. *Journal of Business Strategies*, 18(2), 133-149.
- Filipo, J. P. (2000). *L'innovation dans les activités de service*. Paris: Editions d'Organisation.
- Gallouj, F. (1998). Innovating in reverse: Services and the reverse product cycle. *European Journal of Innovation Management*, 1(3), 123-138.
- IDATE. (2003). *Le marché mondial des services mobiles*. Montpellier, France: IDATE.
- Ishii, K. (2004, February). Internet use via mobile phone in Japan. *Telecommunications Policy*, 28(1), 43-58.
- Lee, M. S. Y., McGoldrick, P. J., Keeling, K. A., & Doherty, J. (2003). Using ZMET to explore barriers to the adoption of 3G mobile banking services. *International Journal of Retail & Distribution Management*, 31(6), 340-348.
- Lehmann, H., Kuhn, J., & Lehner, F. (2004). The future of mobile technology: Findings from a European Delphi study. *Proceedings of the*

37th Hawaii ICSS Conference, Big Island (pp. 1-12).

Loilier, T., & Tellier, A. (1999). *Gestion de l'innovation*. Paris: Edition Management & Société.

MobeyForum. (2003). *MobeyForum mobile financial services*. Retrieved June 2, 2004, from <http://www.mobeyforum.org>

Pousttchi, K., & Schurig, M. (2004). Assessment of today's mobile banking applications from the view of customer requirements. *Proceedings of the 37th Hawaii ICSS Conference*, Big Island (pp. 1-12).

Research and Markets. (2001). *Banking on wireless—European banks blaze the m-bank-*

ing trail. Retrieved June 22, 2004, from <http://www.researchandmarkets.com>

Rogers, E. M. (1983). *Diffusion of innovations* (3rd ed.). New York, The Free Press.

Sangani, K. (2002, March). Mobile banking. *Financial World Magazine*.

Shapiro, C., & Varian, H. R. (1999). *Economie de l'information: Guide stratégique de l'économie des réseaux*. Paris: Edition De Boek Université.

Van de Kar, E., & Van der Duin, P. (2004). Dealing with uncertainties in building scenarios for the development of mobile services. *Proceedings of the 37th Hawaii ICSS Conference*, Big Island (pp. 1-12).

Chapter XXX

Mobile Commerce in Emerging Economies

Amol Patel

ConvergeLabs Corporation, USA

ABSTRACT

This chapter discusses the opportunities and challenges of mobile commerce in emerging economies. It analyses the profound impact of a mobile device on the way products and services are bought and sold in developing nations. The chapter argues that many mobile applications can have a much larger impact on emerging economies than those of the developed world. The chapter is aimed at creating an understanding of the unique social, technological and economic drivers that can help entrepreneurs and solution providers to build and deploy compelling and revolutionary mobile commerce applications in these emerging markets.

INTRODUCTION

This chapter looks at mobile commerce ecosystems in the developing world. It explores the opportunities that manifest themselves in the evolving mobile marketplace, and the underlying value drivers. Furthermore, it also discusses sample mobile commerce applications and the impetus and obstacles to the successful deployment of such mobile commerce applications.

Finally, it considers the future of these mobile applications in the context of the emerging nations.

BACKGROUND

The topics discussed in this chapter are based on the author's entrepreneurial experiences in introducing mobile commerce services in India.

Figure 1. Mobile commerce value chain



During these experiences, the author has had various opportunities to interact with end consumers, mobile operators, and service providers in India as well as other parts of Asia. The resultant knowledge and experiences can be potentially extrapolated to many economies in the developing world, and will hopefully provide invaluable learning opportunities to the reader.

THE MOBILE COMMERCE ECOSYSTEM

The mobile commerce ecosystem consists of the end user, the content and portal, the application developer, the mobile operator, the payment gateway, and infrastructure and device vendors, as shown in Figure 1.

The end user has a mobile device that would have basic voice and data capability. The content is information about the vendor of goods and services (e.g., a movie theatre selling tickets), accessed by the user through a mobile portal. The application developer provides a mobile commerce platform to the mobile operator or directly to the content vendor. The mobile operator provides the wireless network, and markets various voice and data services to its subscribers (i.e., the end users). The payment gateway consists of a clearinghouse for payments via credit/debit cards, bank accounts, and Internet banking. The infrastructure and device vendor provides the wireless equipment for the network and the handsets for the end user.

The following factors affect the mobile commerce value chain: lifestyle of consumers, applications developed to cater to that lifestyle, marketing of those applications, network technology, and security.

LIFESTYLE

People in the emerging economies of Asia spend a lot of time commuting in public transport. They use this “downtime” productively by getting work done through their mobile phone. This has historically led to the rapid adaptation of mobile phones in these markets, as is evidenced in the high levels of ‘SMS’ text messages used in Asia (Infocomm Development Authority of Singapore, 2003) vs. developed western nations like the United States. In contrast to their western counterparts, chances are that consumers in a developing country may never have used (or seen) a computer, and the mobile is their first electronic device that is used for communications (voice and data) and as a “computer”. This convergence provides the impetus to emerging nations to leapfrog the developed world in mobile penetration and usage.

In a country like India, at the time of writing this book, the penetration of mobile phones is already more than 10 times that of the personal computer, and growing rapidly. The number of incremental wireless connections added every month has already exceeded the number of landline ones. Many farmers in rural India are

receiving alerts on their mobile phones about the prices of their produce and making important decisions on their daily sales accordingly. Mobile commerce can provide them the transactional capability to immediately act on the pricing alerts received. Back in 2004, India piloted its first m-ticketing service for selling tickets at an India-Australia cricket match in Bangalore (Shah, 2004), and in a matter of few days, thousands of tickets were sold! End user polls attributed the incredible response from the public to the significant transparency and convenience it provided in obtaining a ticket in a matter of minutes, compared to the long hours one would otherwise have had to stand in line under the scorching tropical sun, and still be unsure of eventually getting the ticket. Cricket is like religion in the world's largest democracy, and your chances of obtaining a cricket match ticket even in a 100,000-plus-capacity stadium are directly proportional to "who" you know in the system, or how much you are willing to spend in the "after-market". It is gratifying to see that mobile technology is leveling the playing field for the masses by bringing in more transparency in ticket purchasing, and the m-ticketing campaign in Bangalore had aptly captured it in its tag line: "The only connection you need to purchase match tickets."

APPLICATIONS

The field of mobile commerce applications is extremely vast, and an entire book can be devoted solely to it. However, there are some applications worth mentioning in the context of developing countries. In many of these countries, the physical infrastructure is very poor. Consequently, people spend a lot of time waiting in lines for routine day-to-day activities like ticketing, retail commerce, banking, and so forth. A mobile platform can provide anytime,

anywhere transactional capability for these mundane tasks. For example, if a user wishes to go to a movie with her friend, she can select the movie and buy the tickets through the mobile portal and receive the tickets directly on her mobile phone as a barcode (i.e., an m-ticket). All she needs to do at the theatre entrance is swipe her mobile on an optical scanner, which automatically scans the m-ticket and authenticates entry (ConvergeLabs, 2004). In case her friend is delayed, she can request separate m-tickets, and have one forwarded to her friend, thereby not missing the beginning of the movie herself! This is an example of selecting, purchasing, and redeeming a ticket entirely through the mobile phone. Similarly, mobile commerce solutions in the business-to-consumer space can provide a mobile storefront to the large variety of retail shops. We have already started seeing basic mobile commerce applications like ringtones and games take off exponentially. Likewise, there are a variety of applications in the business, medical, entertainment, hospitality, and financial sectors which can be enabled through a robust mobile commerce platform (Rutberg & Co., 2004).

MARKETING

The mobile medium provides unprecedented opportunities for marketers to provide customized value propositions to consumers. By leveraging location-based services inherent in mobile networks, *commerce* can occur within a personalized framework of *context*, *content*, and *community*. For example, a user who has 'opted in' can receive an m-coupon "beamed" to her by a nearby McDonald's around lunchtime. Thus a permission based 'pull' model of sales and marketing can be promoted through technology, rather than a traditional "push" (broadcast) model that may give rise to un-

wanted spam. In addition to providing enhanced data mining capabilities for customer relationship management for planned purchases, mobile commerce provides an even greater ability to influence impulse buying through timely messaging. Marketers can develop a one-to-one relationship with their customers anytime and anywhere. We will discuss the various potential business models for monetization of personalized marketing services in a later section of this chapter.

TECHNOLOGY

The average capability of mobile phones prevalent in emerging economies is quite limited, given the low per capita incomes. The infrastructure, in terms of network capabilities and capacities, is also limited, though there are exceptions with advanced 3G networks already in trials in some places in South Asia. For these reasons, hosted mobile commerce platforms are likely to do better than client-server architectures, which need enhanced handset capabilities for client downloads and so forth. Given the large mass of users with low-end phones in places like India, any solution has to cater to the lowest common denominator in terms of handset capabilities.

SECURITY

Security is a crucial aspect of the mobile commerce ecosystem. While technologically the security of mobile commerce solutions may be on par with (or superior to) Internet commerce solutions, the consumer perception too has to be brought to similar levels, through education and awareness. SMS, through which the vast majority of text messaging presently takes place in GSM networks, is not a very secure channel.

Consequently, mobile commerce solution developers are looking for innovative ways of encrypting messaged data, as well as over-the-air downloadable java applications and WAP-based wireless transport layer security solutions.

BUSINESS MODELS

There are different business models that could work in the mobile ecosystem. There could be revenue sharing between the mobile operator and application developer of the “premium” SMS that is charged to users for a particular mobile commerce application. In the movie m-ticket example, a transaction fee per m-ticket can be charged by the application provider. The moviegoer to say a Spiderman movie can also pre-pay for Spiderman toys/t-shirts, which would lead to merchandizing sales commissions revenue sharing within the m-commerce ecosystem. Similarly, the moviegoer can also purchase soda/popcorn in advance (in exchange for say some discount), thereby leading to supply chain management revenue sharing with the concession vendor as well.

Given the low disposable incomes in emerging economies, it is advisable to introduce pricing models that facilitate single-use transactions cost effectively rather than monthly subscriptions. The large volumes in many of these economies would provide for sustainable long-term profit growth once the brand has been established.

We are also entering the age of the mobile professional, who can conduct business anytime, anyplace. This leads to opportunities for enterprise software solutions providers to realize revenue from productivity-enhancing applications like mobile videoconferencing, e-mail, LAN (Local Area Network), and so forth.

PAYMENT MECHANISMS

Internet commerce took off due to the ready online usage of credit cards. In contrast, in many emerging economies, credit card penetration is quite low, but bank account usage is quite high. Hence mobile commerce platforms, in addition to letting users pay for goods and services through credit cards, should also facilitate direct debit of the user's bank account. Mobile operators can also directly charge their subscribers' prepaid or postpaid accounts for the micro-payments done. However, some countries have licensing restrictions that levy surcharges for such transactions by the mobile operator.

IMPEDIMENTS TO GROWTH

There are several potential challenges to mobile commerce growth in emerging economies. Solutions providers that aggressively bring to market half-baked products that have not undergone stringent security testing would discourage consumers from paying from their mobile phones. Location-based services that do not respect the privacy of their users can lead to a backlash. An unchecked 'push' model and consequent spamming can lead to rejection of the service offering. If the price of phones that have the features required to conduct a secure transaction is high, it will dampen the uptake of such services. Quality of service is a serious concern with mobile operators working at near-capacity scenarios, and delays in confirmations of transactions reaching the subscribers can lead to service adoption problems. Mobile operators with a short-term view (like many cable operators in the United States) may be slow to react to the emerging opportunities and try to leverage their monopolistic position in

suppressing innovation and competition. Timely enactment and enforcement of mobile commerce laws are also essential to give the various stakeholders the comfort level needed for mobile commerce.

These are some potential obstacles to widespread adoption of mobile commerce in emerging economies, but hopefully players in the mobile value chain are cognizant of these challenges and also their responsibilities, and will work together to mitigate these risks.

FUTURE DIRECTIONS

The mobile revolution will fundamentally change lives in developing nations much more profoundly and rapidly than the Internet did in the developed nations. This is so because we are sitting merely at the tip of an iceberg in terms of our present understanding of all that mobile commerce can do. There is incredible opportunity for entrepreneurs in emerging economies to find niches and grow those niches into large markets. Some of the most successful companies in the mobile commerce space a few years from now are possibly the ones that have not yet been formed at the present time! It is an exciting time to explore new business models and capture value in dynamically evolving mobile value chains.

REFERENCES

- ConvergeLabs. (2004). *M-Bay: Mobile ticketing and payments solution*. Retrieved August 18, 2005, from www.convergelabs.com
- Infocomm Development Authority of Singapore. (2003, March). *Mobile payments CFC review report*. Retrieved from www.ida.gov.sg

Rutberg & Co. (2004, February). *Research note: 3GSM World Congress* (e-mail newsletter). Retrieved from www.rutbergco.com

Shah, K. (2004, September 17). KSCA to debut mobile ticket to cricket. *The Economic Times (India)*.

Chapter XXXI

M–Business: A Global Perspective

Mahesh S. Raisinghani
Texas Woman's University, USA

ABSTRACT

This chapter discusses the use of mobile, handheld computer devices that are connected wirelessly to a network for business and personal use across people, projects, tasks or organizational units to infer a trend of general acceptance of m-business in the market place. It describes the state of the mobile commerce industry from a worldwide perspective and the barriers to implementation of m-commerce, discusses the issues and challenges followed by the the conclusions and directions for future research.

INTRODUCTION

Only a few years ago, electronic commerce (e-commerce) visionaries were predicting the rapid acceptance of mobile commerce (m-commerce) as the evolutionary result of the e-commerce revolution. M-commerce is the buying and selling of goods and services through wireless handheld devices such as mobile phones and

personal digital assistants (PDAs). The term m-commerce is used in this chapter to describe the adoption and use of mobile, handheld computer devices that are connected wirelessly to a network for business and personal use. Known as next-generation technology, m-commerce enables users to access the Internet without the need to find a place to plug in. According to these visionaries, most of the U.S. population

should be paying bills and shopping with mobile phones or PDAs, while receiving updated flight information on the way to the airport.

In reality, a different situation took place. The wireless industry is lowering its expectations for revenue growth since consumers have not accepted m-commerce as widely as expected. There are signs that m-commerce is growing in popularity. Recently, ComScore Networks (2002) found that almost 10 million wired Internet users in the United States accessed the wireless Internet using mobile devices. Sprint has activated its 3G (third-generation) network in the U.S. market that it serves. Other nations, most notably China, are also moving forward with 3G initiatives. The Chinese are replacing their wireless infrastructures with packet Internet Protocol-based 3G systems in support of the 2008 Olympic Games (Lemon, 2001).

Senior managers will find this information useful for planning and adapting to this new horizon. Educators will also find the information useful as they re-focus academics away from legacy e-business to an m-commerce model in order to adapt themselves to new trends in the world economy. The timing of the shift toward a mobile environment model is important, as strategies must change in advance of the trend. This chapter is structured as follows: we first describe the state of the mobile commerce industry from a worldwide perspective and the barriers to implementation of m-commerce. This is followed by a discussion of issues and challenges. Finally we discuss the conclusions and directions for future research.

Mobile Commerce: State of the Industry

In 2004 we saw the convergence of Wi-Fi and VoIP to evolve a new technology that has been variously described as Voiceover Wi-Fi

(VoWiFi), Voiceover Wireless LAN (VoWLAN), Voiceover Wireless IP (VoWIP), Wireless Voiceover IP (WVoIP), Voiceover IP over Wireless LAN (VoIPoWLAN), Mobile VoIP, and Wi-Fi Telephony. According to a study by Frost and Sullivan Consultancy, about \$25bn (£28bn) in trade will be generated through mobile payments in 2006, or about 15% of estimated online e-commerce consumer spending and mobile-accessed Internet and peer-to-peer payments will make up the bulk of payments, accounting for 39% and 34% of spending in 2006 respectively (M2 Communications, 2002).

The United States, to date tradition-bound by its extensive fixed-line network, is shifting towards wireless means of communication. There are signs that demonstrate an increasing interest by industry towards m-commerce models, but will m-commerce become a reality or will it be just another trend that soon will be outdated? Companies and consumers have fresh memories of past dot-com failures, and this might give one reason to doubt the imminent arrival of a significant level of m-commerce. Furthermore, the high prices of mobile services, together with slow access speeds, have not helped much to add to the luster of the mobile environment. M-commerce in the United States also faces other challenges such as lack of standards, lack of ubiquitous wireless network coverage, technical differences among wireless devices, and security, among others.

Yet despite these drawbacks, mobile wireless devices are a reality today, with expectations of up to 1.5 billion in 2005 (Van Impe, 2002). In 1990, there were five million wireless subscribers in the U.S. In 2000, the number improved to 90 million, and it is expected that by the end of the year 2005, this number will reach 140 million (Kalakota & Robinson, 2001, pp. 26-27). Some expectations go even further, for instance Nokia is predicting that cellular phone

Internet connections will outnumber PC Internet connections by 2004 (Smith, 2002). Leading companies such as Microsoft, Intel, Nokia, NTT DoCoMo, Sony, and Ericsson, among others, are preparing themselves to compete in this arena, taking the first steps that will ensure them their leadership in a future mobile environment.

Europe and most of the Eastern hemisphere have established standards for analog and digital cellular telephone networks. In the United States, however, the Federal Communications Commission (FCC) has set standards for only analog networks, preferring to let each carrier choose its own digital technology. The diversity of these standards (e.g., TDMA, CDMA, and GSM) from vendors means that a given digital handset may not work in all digital service areas, although multimode phones and roaming agreements have alleviated this problem to some extent. In addition, the Czech Republic and Europe in general have one major advantage over the U.S. when it comes to m-commerce: widespread SMS interconnection among mobile operators. Lack of cooperation among operators in the U.S., coupled with a large landmass with limited coverage, has helped Europe take the lead in the race for wireless superiority. SMS in the U.S. has also been hampered by negative pricing policies, since the users have to pay to both send and receive messages (M2 Communications, 2002).

In February 2003, Orange, Telefonica Mobiles, T-Mobile, and Vodafone signed an agreement to form a new Mobile Payment Services Association aimed at delivering an open, commonly branded solution for payments via mobile phones, designed to work across all operator networks. The aim is to make the new initiative available to the largest possible number of mobile phone users. The solution will work across country boundaries and will seek to complement existing industry solutions to be-

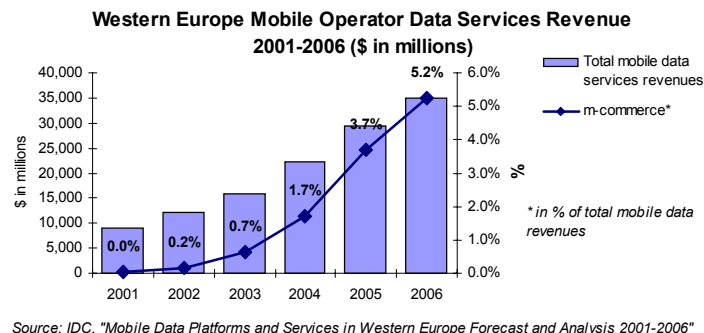
come the industry standard for m-commerce payments (M2 Communications, 2003). For customers, the aim is to provide the opportunity to purchase a wide range of digital and physical goods and services with their mobile phones using an easy, secure solution. Merchants and merchant acquirers will benefit from a standard set of interfaces through which they will gain access to a potentially huge international customer base. Software and solution vendors will benefit from published technical interfaces, enabling the development of compliant m-payment products and services. Operators will benefit from a standard way of integrating and efficiently managing their relationships with merchants, merchant acquirers, and content providers.

In early 2001, Alcatel, Ericsson, Nokia, and Siemens formed the Wireless World Research Forum to guide and promote research of wireless communications beyond 3G. With the deployment of third-generation systems already begun, long-term wireless research and development is focused on fourth-generation technologies, which promise to provide data rates of at least 10 Mbps and more likely 100Mbps; multi-standard capability and seamless interconnections between the multitude of home, workplace, and public wireless networks; quality of service enhancements; and efficient use of bandwidth by assigning frequencies and bandwidths dynamically to users based on services required. Since new-generation wireless phone technologies are introduced approximately once every 10 years, it probably will be 10 years before 4G is launched (PriceWaterhouseCoopers, 2002a).

Figure 1 summarizes the expected growth in m-commerce revenues over the period of 2001-2006.

IDC's forecast shows that total mobile data revenues are expected to be increasing by more than 31% per annum, whereby the CAGR

Figure 1. Western Europe mobile operator data services revenue 2001-2006



for revenues generated by m-commerce is estimated to be more than 265% per annum. As seen in Figure 1, despite rapid growth, income from m-commerce will remain a very small portion of total data revenues (highest value 5.2% in 2006). This implies that in the short run, m-commerce will not turn in profits to justify the investments in the new technologies that were initially believed to boost it. This applies to Europe and the United States. Asia, on the other hand, has developed more quickly with respect to m-commerce. Figure 2 illustrates the mobile operator data services revenue from 2001 to 2006 in Hong Kong and China.

The combined figures for Hong Kong and China show that the total mobile data services revenues are expected to increase for the

period 2001-2006 by more than 70% per annum. The growth in m-commerce revenues in Hong Kong and China is currently 26.3% and is expected to outstrip 45% by 2006. This outstanding acceptance of m-commerce in Asia is only partially due to the currently used mobile technologies that require fewer investments for the upgrade to 3G. The major drivers behind this trend are the habits of the Asians who are keener on innovative technologies, and the variety of content providers that attracts an increasing number of mobile services users.

Interestingly in Asia, the investments were restricted to the lower priced 3G licenses. The service providers, therefore, are able to offer their services at a much lower cost. Especially in China and Hong Kong, where mobile tech-

Figure 2. Hong Kong and China mobile operator data services revenue 2001-2006

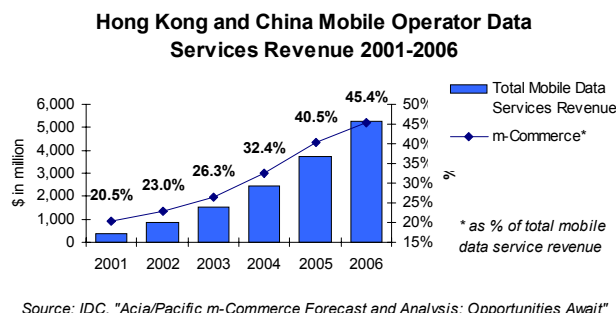
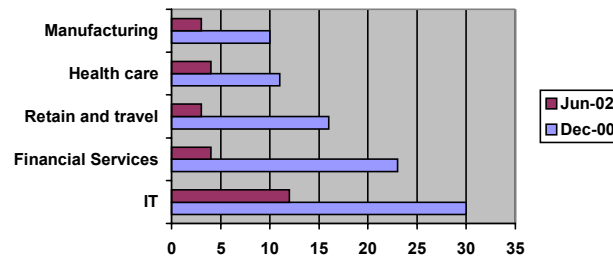


Figure 3. Confidence in wireless e-commerce

Does your company expect that wireless e-commerce will make a major contribution to its e-business revenue within the next 12 months?
(% of respondents answering yes)



Data: Information Week Research E-Business Agenda Study of 372 IT managers, June 2002. Additional data from E-Business Agenda Study, December 2000.

nology was introduced at a later stage and the 2G CDMA standard was adopted initially, the transition to 3G required little investment for the upgrade of the networks. This allows them to offer the services at a lower cost compared to the potential costs to the European consumers. The lower costs make it easier for consumers to try this new technology, become accustomed to and ultimately adopt it.

The mobile market penetration in the U.S. is around 41%, far less than countries like Finland with 75%, Hong Kong with 89%, or the United Kingdom with 74% (Magura, 2003). Besides, only 6% of users in the U.S. use their mobile phones to access the Internet; this is a much lower percentage compared with other countries like Japan with 72%, Germany with 16%, or the United Kingdom with 10% (Beal et al., 2001, p. 6). Business and private consumers in the U.S. use wireless devices mainly to place phone calls, read the news or the weather report, and trade stocks. A popular innovation in the U.S. is close to one million vehicles equipped with satellite tracking and communication devices, most of them with OnStar equipment from GM. OnStar had 800,000 subscribers in the U.S. and Canada in 2001, and ex-

pected four million vehicles equipped with this system by 2003 (Kalakota & Robinson, 2001, p. 6). Still, 61% of U.S. consumers think that wireless Internet access is too expensive, among other drawbacks including difficulty to read the screen (61%), slow access (37%), privacy concerns (16%), and unreliable service (12%) (Beal et al., 2001, p. 50).

Despite momentum in the consumer markets for m-commerce, many in the business community still remain pessimistic about its future for the enterprise. Consider Figure 3 which details the more pessimistic attitude prevailing in the U.S. marketplace for consumer applications.

Interestingly, there is an inverse relationship between pessimism and size of the company. Overall almost 75% of small companies did not expect m-commerce revenue, but only 58% of large companies felt the same way (Ewarl, 2002).

Barriers to Implementation

Among the key challenges in mobile and wireless information systems are issues related to internetworking and integration of different

wireless technologies, reliable and secure communications, context and location-awareness in both applications and mobile devices, device and user interface issues, accessibility of wireless networks for continued business operation and support for group communication by allowing multicasting in the network, and application and/or middleware protocols (Varshney, 2003). For instance, the issue of PDA/cell phone and screen size, with respect to the amount of information that can be presented on such a small display, or infrastructure issues and problems of access, coverage, roaming, reliability, location management, and multicast communications are problems that are currently being addressed in the research community. Carriers will need to spend large sums to deploy sufficient networks, and the key question with the current state of the industry is: how much more money are telecommunications carriers likely to spend to ensure availability and performance of the carrier network and sufficient coverage for data transmission? Additionally, interoperability issues are of concern due to the lack of common standards (Baldwin, 2002). Many wireless networks are filled to capacity. For instance, AT&T struggles to keep the number of blocked calls under six million a day, not including the calls that are connected and then dropped, with spotty coverage increasing attrition among customers (Woolley, 2002).

Another concern regarding carriers is that there are three different and incompatible technologies deployed by U.S. wireless telecom companies that have resulted in three different paths to 3G. Sprint can offer 3G more rapidly and more cost effectively than AT&T because their wireless network is compatible, while AT&T would have to build a new network and was not able to launch 3G until 2003. Cingular is not planning a 3G roll-out and believes that demand will be weak, but other industry ana-

lysts believe Cingular lacks sufficient spectrum, and parents SBC Communications and BellSouth each use different types of wireless technologies, making a 3G solution much more expensive (Lanners, 2002).

Anecdotal evidence revealed that while there is a great deal of optimism for the eventual arrival of m-commerce, network infrastructure dominated as the most critical barrier to diffusion of m-commerce. Wareham and Levy (2002) provide additional insight in their study on mobile telecom diffusion and possible adopters of 3G computing devices. They noted that early adopters of wireless telephone devices correlated positively with executive-level positions and income levels. Of further significance to the authors was the study's conclusion that adoption of m-commerce applications would be triggered with the introduction of a 3G network into the marketplace.

ISSUES AND CHALLENGES

A familiar problem many companies face today is accessibility. Employees who work in areas where there is no cellular coverage cannot be reached. This adversely affects productivity. VoWLAN provides a solution to this problem. A company can setup an enterprise-wide Wi-Fi network used to deploy VOIP wirelessly. Successful deployment requires a comprehensive wireless coverage to enable all users in the organization to be connected at all times.

Unlike data traffic, voice traffic is very sensitive to connection and latency issues. A thorough site survey is required to plan a good deployment strategy. The latency factor requires the voice traffic to have uncontained connections. Thus, manufacturers specify and limit the number of connections at an access point to seven simultaneous connections, even

though VoWLAN does not require a very high bandwidth. However, solutions are available with deployment strategies which can accommodate up to 30 connections. 'ON World' projects a market growth of 83% for VoWLAN handsets by 2007 and a Wi-Fi equipment market growth of 40% for the next five years. Newer technologies and standards in Wi-Fi and VOIP are emerging which will open doors to a whole array of applications for VoWLAN. With the development of standards like WiMAX and SIP and standardization of VOIP, cellular vendors are recognizing VoWLAN as a value-added feature, rather than a competitor, to their cellular phones.

Customers will use VoWLAN and cellular technologies interchangeably. When they work in Wi-Fi hotspots, they will take advantage of VoWLAN and make low-cost calls worldwide. Outside the Wi-Fi envelope, they can take advantage of the mobility and coverage of traditional cell phone technologies. Today, Wi-Fi is a dynamic and growing market, establishing its presence in both homes and enterprises. Wireless-enabled devices like laptops, handhelds, and even cellular phones are becoming increasingly popular, with more than 850 million Wi-Fi-enabled cell phones expected to be shipped by the year 2009.

VoWLAN is a new era in the evolution of personal mobile communication. Market potential and emerging technologies have driven many companies to plunge into the VoWi-Fi wave. Among the leading vendors are Cisco, Nortel Networks, SpectraLink, Symbol Technologies, Telesym, Texas Instruments, and Vocera. The huge market, converging technologies, emerging technologies, and value-added application are the driving forces for VoWLAN to be considered as the solution of the future.

CONCLUSION AND DIRECTIONS FOR FUTURE RESEARCH

Although m-commerce in the short term is facing some difficulties, companies have invested heavily into a range of technologies that could facilitate m-commerce. These technologies, especially the 3G, are able to offer bandwidth and speed that customers have become accustomed to through land-wired access. Unlike e-commerce, however, industry has not been able to adapt one common standard. These standards are necessary for the scalability and reach of services offered to the customer. Pricing, coverage, and functionality are the key drivers for consumers in the United States to adopt new wireless technologies and services. Without products and services that add value for the customer, users will not take to m-commerce. At this point, customers have not become acquainted with the technology and have not gained trust in the security options offered. To overcome these obstacles, intensive efforts by the service providers, content providers, and retailers are required to market m-business. In addition, a concerted effort is necessary in order to agree on a worldwide standard that will eventually trigger a wider variety of content offered. In the long term, customers are more likely to change their behavior and use services that offer an improvement to their lifestyle by saving time and/or adding convenience.

REFERENCES

Baldwin, H. (2002, September 20). The challenges of mobile development. *ZDNet*. Retrieved from <http://techupdate.zdnet.com/techupdate/stories/main/0,14179,2881040-1,00.html>

- Basso, M. (2001, December 5). Will mobile portal's future be bright? *Gartner Dataquest*. Note AV-14-9502.
- Beal, A., Beck, J. C., Keating, S. T., Lynch, P. D., Tu, L., Wade, M., & Wilson, J. (2001, June 4). *The future of wireless: Different than you think, bolder than you imagine*. Retrieved from http://www.accenture.com/xd/xd.asp?it=enWeb&xdn=_isc/iscresearchreportabstract_134.xml
- Clarke, R. (1999, September 26). *A primer in diffusion of innovation theory*. Retrieved from <http://www.anu.edu.au/people/Roger.Clarke/SOS/InnDiff.html>
- ComScore Networks. (2002, August 27). *Ten million Internet users go online via a cell phone or PDA*. Retrieved from http://www.comscore.com/news/cell_pda_082802.htm
- Downing, C. E. (1999). Systems usage behavior as a proxy for user satisfaction: An empirical investigation. *Information and Management*, 35(4), 203-216.
- Evans, N. (n.d.). *The m-business evolution: Business agility: Strategies for gaining competitive advantage through mobile business solutions*. Retrieved from <http://www.developer.com/ws/other/article.php/1446771>
- Fichman, R. G. (2001). The role of aggregation in the measurement of IT-related organizational innovation. *Management Information Systems Quarterly*, 23(4), 4.
- Harman, H. H. (1976). *Modern factor analysis*. Chicago: The University of Chicago Press.
- Kalakota, R., & Robinson, M. (2001, October). *M-commerce: The race to mobility*. New York: McGraw-Hill.
- Kort, T. (2002, January 21). Handheld market to grow 19% in 2002. *Gartner Dataquest*. Note HARD-WW-DP-0194.
- Kwon, T., & Zmud, R. (1987). Unifying the fragmented models of information systems implementation. In R. J. Boland & R. A. Hirschheim (Eds.), *Critical issues in information systems research* (pp. 87-93). New York: John Wiley & Sons.
- Lanners, J. (2002). Interview. *Sprint*, (October 1).
- Lemon, S. (2001, July 17). Beijing's Olympic success spurs 3G vision. *IDG*. Retrieved from <http://www.nwfusion.com/news/2001/0719olympic.html>
- Levin, R. I., & Rubing, D. S. (1994). *Statistics for management* (6th ed.). Englewood Cliffs, NJ: Prentice-Hall.
- M2 Communications. (2002, May 31). Examining m-commerce in Central Europe. *Europemedia*, p. 1.
- M2 Communications. (2003, February 25). New m-commerce association formed. *Europemedia*, p. 1.
- Magura, B. (2003, Spring). What hooks m-commerce customers? *MIT Sloan Management Review*, 44(3), 9.
- Mak, B. L., & Sockel, A. (2001). A confirmatory factor analysis of IS employee motivation and retention. *Information and Management*, 38(5), 265-276.
- Mertz, C., & Serrell, M. D. (2002, October 15). Mobile application tools. *PC Magazine Online*. Retrieved from <http://www.pcmag.com/article2/0,4149,545121,00.asp>
- Moore, G. C., & Benbasat, I. (1991). Development of an instrument to measure the perceptions of adopting an information technology innovation. *Information Systems Research*, 2(3), 192-222.
- Nobel, C. (2001, March 19). IT is lukewarm to wireless consumer apps. *eWeek*. Retrieved

- from <http://www.eweek.com/article2/0,3959,104226,00.asp>
- Nunnally, J. C. (1978). *Psychometric theory*. New York: McGraw Hill.
- PriceWaterhouseCoopers. (2002a). *Technology forecast: 2002-2004*, 2, 453.
- PriceWaterhouseCoopers. (2002b). *Technology forecast: 2002-2004*, 2, 680.
- Rai, A., & Howard, G. S. (1994). Propagating CASE usage for software development: An empirical investigation of key organizational correlates. *Omega*, 22(2), 133-147.
- Rogers, E. M. (1983). *Diffusion of innovations*. New York: The Free Press.
- Rogers, E. M. (1995). *Diffusion of innovations* (4th ed.). New York: The Free Press.
- Sharma, S., & Rai, A. (2003). An assessment of the relationship between ISD leadership characteristics and IS innovation adoption in organizations. *Information and Management*, 40(5), 391-401.
- Smith, D. (2002). *The pocket computer*. Retrieved from <http://www.ed2go.com/news/wireless.html>
- Tornatzky, L. G., & Klein, L. (1982). Innovation, characteristics and innovation-implementation: A meta-analysis of findings. *IEEE Transactions on Engineering Management*, 29(1), 28-45.
- Uncapher, M. (2002, March). *Mobile commerce: WITSA/WIRG survey*. Retrieved from <http://www.itaa.org/isec/pubs/e20023-04.pdf>
- Van Impe, M. (2002, June 19). *Nokia expects the number of mobile users to surge in the next three years*. Retrieved from http://www.mobile.commerce.net/story.php?story_id=1824
- Varshney, U. (2003). Location management for mobile commerce applications in wireless Internet environment. *ACM Transactions on Internet Technology*, 3(3), 236-255. Retrieved from <http://ftp.informatik.unitrier.de/~ley/db/journals/toit/toit3.html#Varshney03>
- Wareham, J., & Levy, A. (2002). Who will be the adopters of 3G mobile computing devices? A profit estimation of mobile telecom diffusion. *Journal of Organizational Computing and Electronic Commerce*, 2(2), 162-174.
- Woolley, S. (2002, May 27). Zeglis the Zealot. *Forbes*, p. 58.

Chapter XXXII

Mobile Strategy Roadmap

Francesco Falcone

Digital Business, Italy

Marco Garito

Digital Business, Italy

ABSTRACT

Convergence between business environment and technology solutions is a today challenge: is it possible to identify and adapt traditional business analysis tools into IT infrastructure and viceversa ? Mobile business has already dramatically changed our way of life: to get the most of it, it is necessary to understand how and when to take the next step in order to achieve the best possible results. After a short description of the current and available Mobile technologies, the chapter tackles the solutions that some companies have already implemented, enabling thus customers and stakeholder to efficiently and effectively cooperate: yes, because the big news in mobile business is that such a technology makes business process and communications (either internal and external) visible. It is a true “one stop shop” where people are really always connected. It is now possible to create a framework where business environment and mobile solutions get together, enabling the development of a roadmap (or a mindset if you prefer so) that can be used to create new services. Enjoy the reading!

INTRODUCTION

This chapter outlines the convergences and the opportunities available in mobile business markets from both technology and business perspectives. An overview of the wireless options available to businesses is provided. This is followed by description as well as analysis of the mobile business in order to provide a possible framework enabling design, development,

and eventual implementation of new business initiatives incorporating mobility.

THE MOBILE BUSINESS ENVIRONMENT

Mobile business is a broad definition that includes communication, transactions, and different valued-added services that are made

available by using portable devices. Today, most of the attention is around consumer services, but business-to-business and business-to-employee segments are important too.

Another essential definition of “mobile commerce” is that it is referred to as “transactions with monetary value, conducted by mobile Internet.” This definition covers business-to-business, business-to-consumer, and consumer-to-consumer transactions. Traditional voice calls are not included in the definition of mobile commerce, but the services using voice recognition systems to enable commercial transactions fall into the category. Mobile commerce is a subset of electronic commerce in terms of technical issues. However, the term “mobile e-commerce” is misleading because the business models and the value chain of mobility differ from electronic commerce.

Mobile commerce is not a truncated form of e-commerce, but an innovative way of conducting time-critical transactions regardless of location (May, 2001; Paavilainen, 2002).

Wirelessly Speaking

The wireless world is a complex environment consisting of different and competing technologies. Some of these technologies are as follows:

- **WiFi:** The first WLAN (wireless local area network) standard offering the capability to connect wirelessly to LAN; this technology developed rapidly with a wider offer of bandwidth (Gratton & Gratton, 2004).
 - **Strengths:** Expanding existing network without cables; expanding network where cables are difficult to install, rapidly evolving with users’ need; largely used and close at hand; one of the fastest wireless technologies available; very flexible for home and small business
- **Weakness:** Difficult set up and configuration
- **Competitors:** HiperLAN
- **Complements:** Bluetooth, WirelessUSB, and ZigBee
- **HiperLAN:** This technology is used in Europe and provides a different set of wireless communication specifications providing WLAN support. It is also compatible with 3G, enabling voice, and imaging communications (Gratton & Gratton, 2004).
 - **Strengths:** WLAN support provided, data rate of 54 Mbps, the 3G applications support
 - **Weakness:** Located only in Europe
 - **Competition:** WiFi
 - **Complements:** Bluetooth, WirelessUSB, and ZigBee
- **Bluetooth:** From the name of the Viking king Harald whose aim was to join together the Nordic European territories, a cable replacement technology to overtake the messy cables around laptops and desktops. Infrared was also engineered to tackle similar issues (Gratton & Gratton, 2004).
 - **Strengths:** Can be incorporated in many products and devices, low cost resulting in cheaper products, can make use of low power schemes, ease of use for consumer
 - **Weaknesses:** Small data rate available, uses the same frequency as other devices
 - **Competition:** WirelessUSB, some overlap with ZigBee
 - **Complements:** HiperLAN, WiFi
- **ZigBee:** Introduced to disseminate a large range of products and devoted to businesses willing to develop wireless prod-

ucts. It works with IEEE to set up a new standard (Gratton & Gratton, 2004).

- **Strengths:** Low power consumption and cost, affordable wireless solutions, fast wireless development, avoids coexistence issues
- **Weakness:** Uses the same frequency of other technologies and may overlap with Bluetooth
- **Competition:** WirelessUSB and some overlap with Bluetooth
- **Complements:** HiperLAN, WiFi
- **WirelessUSB:** A short-range wireless technology operating on the 2.4 GHz unlicensed spectrum; developed by Cypress Semiconductors to gap fill Bluetooth shortcomings (Gratton & Gratton, 2004).
 - **Strengths:** Low power consumption, affordable wireless solutions, allows fast wireless development
 - **Weaknesses:** There are restricted applications, uses same frequency of other technology, mostly unknown and not widely available
 - **Competition:** Bluetooth and overlap with ZigBee
 - **Complements:** WiFi, HiperLAN
- **Ultra Wide Band:** The FCC in the U.S. recently approved this technology, which is similar to Bluetooth, but can travel a distance up to 230 feet through obstacles by using minimal power. Currently two types of application exist: radar and voice/data communications.
 - **Strengths:** Low power consumption, can overtake obstacles, its radio waves travel further
 - **Weakness:** Early development stage and mostly unknown
 - **Competition:** HiperLAN
 - **Complements:** None yet (Gratton & Gratton, 2004)

- **WiMAX:** Represents the next evolution in broadband wireless technology and will be backed by Intel. This new technology, designed so that it does not require line of sight, should allow higher-speed downloads over much longer ranges than WiFi. In part this is because devices will support certain licensed spectrum bands, enabling them to transmit at higher power levels. WiMAX should have clear advantages of speed and simplicity over 3G technologies for in-vehicle entertainment, flexible CCTV, and security systems; WiMAX devices could represent a user's second or third broadband connection

Does it make sense to have a converging strategy that combines fixed and wireless Internet business? The answer is yes, at least for DHL, the parcel delivery company: the fixed Internet tracking system took six months to be fully available, while the WAP version for the same service was made in seven days (Ahonen, 2002). What does it tell us? It is telling us that one of the success factors (with direct impact on bottom line and time constraints) is the reusability of technology, and it seems that converging technologies could deliver this effect.

Table 1 shows the advantages and disadvantages between a fixed Internet and Mobile Internet, combining the business side and behavioral side.

We personally recommend bearing in mind the above synopsis as it will be very helpful later. It is now time to take a closer look at what the current business scenario is in the mobile business: we are going to provide some examples encompassing at the same time general and common services and more specific industry-focused initiatives carried out in several companies.

Table 1. Advantages and disadvantages of fixed vs. Mobile Internet

ISSUES	FIXED	MOBILE
NUMBER OF USERS	Less	More
COST OF SERVICE	More because of infrastructure	Less
BILLING	Difficult to determine and many think it should be free	Easy: people are aware that making a call has a cost
MICRO-PAYMENT	Very difficult as commission fee for transactions, set up by credit card issuers, are or can be expensive	“Incorporated” in the mobile device and into the business model behind it

The Wireless World Today

We are going to take you into the world of m-commerce services with some simple descriptions of very promising applications. Let us start with *location-based services (LBSs)*: these are an example of how service providers can make use of the inherent properties of mobile devices. Users are always carrying their devices with them and the mobile operator can localize the device, therefore providers can localize them with the help of the mobile operator (Ahonen, 2002; Eurescom, 2004).

Micro-Payment

Payment is an important issue when it comes to adoption and acceptance of services by customers. The customer can turn his mobile phone into a payment device and use it to pay for items and services at a real or virtual point of sale. A micro-payment example of converging technology is given by Coca-Cola vending machines: in Finland users can send an SMS message to the vending machine and pay through their phone bill; the same opportunity is available in the U.S., Poland, Australia, and Hong Kong (Ahonen, 2002, 2003). This is a confirmation that mobile business and micro-payments work well together, that people are available to use mobile phones to make small purchase of

goods and services. What about the business side? One of the strengths of Coca Cola is distribution (business side), while one of the strengths of a mobile phone is portability (consumer side): two sides of the same coin. The parallel lesson is that the same services or applications for either business or consumers can be easily and cost effectively reciprocated.

Gambling

Gambling and betting services, such as lotto, instant games, and sports betting, are very popular in the real world. “M-users” can place their bets using text-based technologies like SMS or WAP, or they can play games. Gambling is an excellent example of how an entertainment service could attract customers by offering a rich, though often mainly text-based contest, with a degree of user interactivity and a real-time user experience (Ahonen, 2002; Eurescom, 2004).

Intelligent Advertising

The basic idea of an intelligent advertising service is that customers receive advertisements (e.g., via SMS or MMS) from merchants on their mobile phones that are adapted to their personal preferences and location based (this is an example of how to combine two different

services, thus providing a possibly unique experience for customers). The service requires a close collaboration between service providers (later on, some additional information about this point will be provided). The customer has to give permission (opt-in approach) in advance to receive ads from the mobile operator. The main advantage of the service is that merchants can be sure to reach the right person by knowing the user profile, so they are likely to be willing to pay for it; therefore, from a marketing point of view, it is possible to estimate how many prospects are in a given marketplace, the cost to get to them, and the ROI (return on investment) ratio (for more about the opt-in approach, see Roman & Hornstein, 2004).

It is time now to take a broader look at the wireless world as it is now.

Current Mobile Services Available in the Business Environment

New technology seems to suggest that mobile services will be the greatest opportunity for businesses to develop richer and more profitable relationships with individual customers by giving them what they actually want—once more a confirmation that wireless technology is an enabling tool. But how is business spelling this definition? Let us have a look at some examples.

Keebler Co., a subsidiary of The Kellogg Co., is the first consumer packaged goods company to use instant messaging to enhance its ties with customers of company brands: the service is called “RecipeBuddie” and allows customer to get recipes based on their mood or food preferences. The service is available for AOL and MSN platforms: the service starts as soon as the customers send an instant message to the screen name “RecipeBuddie” (Newell, 2004).

Another example comes from Land’s End: the company launched a new service named Land’s End Live, enabling customers to interact with service representatives; the service has been enriched by a new service called “Shop with a Friend,” where two customers can exchange messages while shopping (Newell, 2004).

In the United Kingdom, Safeway, one of the largest retail companies, gives customers a Palm-powered device with small magnets on the back of them, so customers can put them on their fridge: when a product is finished, the customer checks off the product and brand, and a request is transmitted wirelessly to the nearest Safeway shop (Newell, 2004).

Now, just imagine a motor insurance that is calculated on how often, where, and when someone is driving: this is the approach of Norwich Union, the largest UK insurer and part of Aviva PLC. The service is called “Pay As You Drive”: each customer’s car has a black box to record location and time of each trip made by the car. The data is sent to Orange network, where the basic information is mapped to a broad network database to enhance the data with the road types and numbers. Premiums are calculated accordingly, and customers can check the sum month by month, so they can change their driving patterns or behaviors. Norwich Union is thus able to adapt its value offer by matching customers’ lifestyle by providing more than insurance products (check the company Web site for the latest deal).

Again in the U.S., Fidelity launched a wireless service based on the success of its Instant Broker initiative, called Fidelity Anywhere, which enables customers to manage their financial position; soon other services will be included (Newell, 2004; Harris & Dennis, 2002).

Abbey National offers e-banking services through the Genie mobile portal, which can be

reached by WAP cellular phones by using any network (Newell, 2004; Harris & Dennis, 2002).

United Airlines in 1999 offered Palm-based service and now gives its customers the capability to be notified when flight status changes and a WAP-based service for last-minute updates and domestic booking for frequent flyers. Similar services are currently available in some hotel chains such as Holiday Inn, InterContinental, and Six Continents Products like i-SPOT Personal Item Locator by Digital Innovation and FINDIT, an electronic locator made by Ambitious Ideas, are already available. Similar scenarios are provided by RFIDs (radio frequency identification tags) with the smart management of supply chain, stock items, and procurement, now possible in real-time mode (Newell, 2004; Harris & Dennis, 2002).

So far we have seen mobile business as a part of the fixed Internet: can we imagine a use of wireless with other media, achieving a converging environment with a unique user experience? Of course we can.

A Set of Converging Technology Examples

In Germany, RTL-teletext, which offers a TV broadcast message board, claims to host up to 220,000 text messages a day. Almost 70% of the broadcasters in Europe have now launched their own SMS chat lines and enjoy similar success: SMS provides excellent indications of a show's popularity or potential, even though conversion rates (ratio viewer/SMS participant) vary widely by application and by the content of individual shows. McKinsey (Bughin, 2004) demonstrates that if more than 5% of a show's viewers interact with it, its audience is extremely engaged and more available in referral and word of mouth.

Viewers who use SMS-TV to vote for contestants on the hit Big Brother, for example, buy

more show-related merchandise than do other viewers, and 70% of the teenagers who purchase Big Brother merchandise vote by text message, according to McKinsey research (Bughin, 2004). SMS interactivity can encourage ratings growth of 50 to 100% for niche cable and satellite channels. Advertising provides 20% of the revenues of the average thematic pay broadcaster, which can reasonably expect one out of every five shows to be interactive. A standard TV show is produced and packaged in an appropriate length: an SMS-TV (basically, sending an SMS to a given number displayed during a TV show) extends the lifecycle of the show. Let us see which are the key success factors.

- **SMS Must be Well Integrated and Synchronized with the Show's Content.**
- **A Show Host with "Push":** The host's suggestions should always be timely and subtle; viewers do not want to feel obligated.
- **Clear SMS Displays:** On-screen numbers and SMS displays are large and clear.
- **Sufficient Rewards:** Tangible rewards (right combination of online and off-line environment).
- **SMS Interactivity:** The broadcaster's production and technical departments should work together, under the same roof, and test the planned SMS interactivity (never underestimate people and process issues).

But how did the subject matter experts spell the words "mobile business"? NTT DoCoMo used the acronym "MAGIC," which stands for Mobile, Anytime, Globally, Integrated, and Customized. Ericsson developed the "0-1-2-3" approach: 0 written manuals, 1 simple button to the Internet, 2 seconds of delay waiting to

Table 2. Communication strategy (Luftman, 2002)

	STEP 1 AD HOC PROCESS	STEP 2 COMMITTED PROCESSES	STEP 3 ESTABLISHED PROCESS	STEP 4 IMPROVED PROCESS	STEP 5 OPTIMIZED PROCESS
UNDERSTANDING OF BUSINESS BY IT	IT management lacks understanding	Limited understanding by IT managers	Good understanding by IT management	Understanding encouraged among IT staff	Understanding required at all levels
UNDERSTANDING OF IT BY BUSINESS	Managers lack understanding	Limited understanding	Good understanding by managers	Understanding encouraged among staff	Understanding required of staff
ORGANIZATIONAL LEARNING	Casual conversations and meetings	Newsletters, reports, and group e-mail	Training, departmental meeting	Formal methods sponsored by senior management	Learning monitored effectiveness
STYLE AND EASE OF ACCESS	Business to IT only: formal	One way, informal	Two way, formal	Two way, sometimes informal	Two way, informal and flexible
LEVERAGING INTELLECTUAL ASSETS	Ad hoc	Some structured sharing emerging	Structured around processes	Formal sharing across the board	Formal sharing with partners
IT/BUSINESS LIAISON STAFF	None or use if needed	Primary IT/business link	Facilitate knowledge transfer	Facilitate relationship building	Build relationships with partners

Table 3. Metrics development (Luftman, 2002)

	STEP 1 AD HOC PROCESS	STEP 2 COMMITTED PROCESSES	STEP 3 ESTABLISHED PROCESS	STEP 4 IMPROVED PROCESS	STEP 5 OPTIMIZED PROCESS
IT METRICS	Technical only	Technical costs, metrics rarely reviewed	Review, act on technical ROI metrics	Also measure effectiveness	Also measure business opportunities, HR, partners
BUSINESS METRICS	IT investment rarely measured	Cost/unit rarely reviewed	Review, act on ROI, cost	Also measure customer value	Balanced scorecards, includes partners
LINK BETWEEN IT AND BUSINESS METRICS	Value of IT investments rarely measured	Business and IT metrics not linked	IT and business linked	Formally linked, reviewed, and acted upon	Balanced scorecards, includes scorecards
SERVICE-LEVEL AGREEMENTS	Use sporadically	With units for technology performance	With units, becoming enterprise wide	Enterprise wide	Includes partners
BENCHMARKING	Almost never	Sometimes informal	May benchmark formally, seldom act	Routinely benchmark, usually act	Routinely benchmark, act, and measure results
FORMALLY ASSESS IT INVESTMENTS	No assessment	Only when there is a problem	Routine occurrence	Routinely assess and act on findings	Routinely benchmark, act, and measure results
ONGOING IMPROVEMENT PRACTICES	None	Few, effectiveness not measured	Few, starting measuring effectiveness	Many, frequent assessments	Practices and measures well established

Table 4. Governance and policy (Luftman, 2002)

	STEP 1 AD HOC PROCESS	STEP 2 COMMITTED PROCESSES	STEP 3 ESTABLISHED PROCESS	STEP 4 IMPROVED PROCESS	STEP 5 OPTIMIZED PROCESS
FORMAL BUSINESS STRATEGY PLANNING	Undone or done when needed	At unit functional level, slight IT input	Some IT input and cross-functional planning	At unit enterprise, with IT	With IT and partners
FORMAL IT STRATEGY PLANNING	Undone or done when needed	At unit level, slight business input	Some business input and cross-functional planning	At unit enterprise, with business	With partners
ORGANIZATION STRUCTURE	Centralized or decentralized	Centralized or decentralized, some co-location	Centralized, decentralized, or federal	Federal	Federal
REPORTING RELATIONSHIP	CIO reports to CFO	CIO reports to CFO	CIO reports to COO	CIO reports to COO or CEO	CIO reports to CEO
HOW IT IS BUDGETED	Cost center, spending is unpredictable	Cost center by unit	Some projects considered as investments	IT = investment	Profit center
RATIONALE FOR IT SPENDING	Reduce cost	Productivity, efficiency	Also a process enabler	Process driver, strategy enabler	Competitive advantage, profit
SENIOR-LEVEL IT STEERING COMMITTEE	Does not exist	Meet informally as needed	Formal committee meet regularly	Proven to be effective	Also with external partners
PRIORITIZATION OF PROJECTS METHOD	Upon IT or business need	Determined by IT function	Determined by business function	Mutually determined	Partners' priorities included

access the service, and 3 keys to gain access to services and features (Ahonen, 2002).

What is a possible approach to develop a strategy for the next wave, without possibly reinventing the wheel?

MOBILE BUSINESS ROADMAP TO SUCCESS

In our opinion, it is necessary to have well clear in mind Porter's 5 Forces Analysis (Porter, 1980) and the SWOT (strengths, weaknesses, opportunities, and threats) scheme, and do the assessment twice: firstly at higher level (macro-economic level) and secondly at lower/specific industry level (micro-economic). This will help

to determine the positioning and the market(ing) potential. This is nothing new: get big, get a niche, or get out (or do not enter at all). Is there a way to comprehend mobile business with its specific characteristics to create a tool enabling us to developing successful business initiatives? This is what the following considerations will try to do.

A preliminary issue needs to be solved before talking about mobile business: the alignment between Business and Technology and the following tables explain how to deal with the internal organization.

By matching rows and columns, it is possible for you to see what the situation looks like—what are the challenges and what is a foreseeable way out. Once you are done with the

Table 5. Partnership decision process (Luftman, 2002)

	STEP 1 AD HOC PROCESS	STEP 2 COMMITTED PROCESSES	STEP 3 ESTABLISHED PROCESS	STEP 4 IMPROVED PROCESS	STEP 5 OPTIMIZED PROCESS
BUSINESS PERCEPTION OF IT	Cost of doing process	Becoming an asset	Enabler of future activities	Driver of future activities	Partner with business of creating value
IT ROLE IN STRATEGIC BUSINESS PLANNING	Not involved	Enables business processes	Driver of business processes	Enabler or driver of business strategy	IT/business adapt quickly to change
SHARED RISKS AND REWARDS	IT takes the risks, no rewards	IT takes the most part of the risks, little reward	IT and business start sharing risks and rewards	Risks, reward always shared	Managers get incentives to take risks
MANAGING THE IT/BUSINESS RELATIONSHIP	Not managed	Managed upon need	Processes exist but not always followed	Processes exist and complied with	Ongoing improvement of processes
RELATIONSHIP/TRUST STYLE	Conflict and mistrust	Transactional relationship	IT as a valued service provider	Long-term relationship	Partner, trusted vendor of IT services
BUSINESS SPONSOR/CHAMPIONS	None	Other have a senior IT sponsor/champi on	IT and business champion at unit level	Business sponsor/cham pion at corporate level	CEO is the business sponsor

Table 6. Technology implementation (Luftman, 2002)

	STEP 1 AD HOC PROCESS	STEP 2 COMMITTED PROCESSES	STEP 3 ESTABLISHED PROCESS	STEP 4 IMPROVED PROCESS	STEP 5 OPTIMIZED PROCESS
PRIMARY SYSTEMS	Office support	Transactional oriented	Business process enabler	Business process driver	Business strategy enabler/driver
STANDARDS	None or not enforced	Defined and enforced at functional level	Emerging coordination across the functions	Defined and enforced across functions	Also coordinated with partners
ARCHITECTURAL INTEGRATION	Not well integrated	Inside unit	Integrated across functions	Begins to be integrated with partners	Fully integrated
HOW IT INFRASTRUCTUR E IS INTEGRATED	Utility, run at minimum cost	Start to be driven by business strategy	Driven by business strategy	Helps business to respond to change	Enables fast response to changing market

previous exercise, then you can take the next step.

Five Lenses to See Mobile

Mobile business can be analyzed into five different wedges: movement, moment, me, money,

and machines. The first one, *movement*, is the most obvious, as movement or changing places is a natural characteristic of mobile phones and their networks. Mobility includes further concepts such as mobility, locality, global, home base, and positioning (Ahonen, 2002).

Mobile Strategy Roadmap

Table 7. Human resources (Luftman, 2002)

	STEP 1 AD HOC PROCESS	STEP 2 COMMITTED PROCESSES	STEP 3 ESTABLISHED PROCESS	STEP 4 IMPROVED PROCESS	STEP 5 OPTIMIZED PROCESS
INNOVATIVE AND ENTREPRENEURIAL ENVIRONMENT	Discouraged	Sometimes encouraged at unit level	Strongly encouraged at unit level	Also at corporate level	Also with partners
KEY IT HR DECISION MAKERS	Top business and IT management at corporate level	Same, with emerging functional influence	Top business and unit management, IT advises	Top business and IT management across firm	Top management across firm and partners
CHANGE READINESS	Tend to resist change	Change readiness programs emerging	Programs in place at functional level	Programs in place at corporate level	Proactive approach to anticipate the change
CAREER- CROSSOVER OPPORTUNITIES	Rare job transfers	Occasional transfer inside the unit	Regularly happens for unit management	Regularly happens at all unit levels	Also at corporate level
CROSS- FUNCTIONAL TRAINING AND JOB ROTATION	No opportunities	Decided by units	Formal program run by all units	Also across the enterprise	Also with partners
SOCIAL INTERACTION	Minimal IT/business interaction	Only business relationship	Trust and confidence start	Trust and confidence achieved	Fully achieved with partners and customers
ATTRACT AND RETAIN TOP TALENTS	No retention, poor recruiting	IT hiring focused on technical skills	Technology and business focus, retention programs	Formal programs for retention and recruiting	Effective programs for retention and hiring

The service can transfer with the user as long as the user moves; the changing user pattern highlights the need to have an easy-to-use device and be in contact with one's own personal relationships, get information and data when a need should arise, in one word: communicate. Therefore, the enabled device must be easy to use (the above mentioned 0-1-2-3 or Magic approach), portable (and current models of mobile phones bear this feature), and have the capability to perform simultaneous—convergent—purposes (not only phone calls, but also messages, pictures, data, or games, just to name a few). Moreover, mobility can assume concurrent faces depending on location either narrowly (time and map of a specific place) or globally (when someone wants to be informed about stock exchanges). The main task for developing and providing a smart service is to be user sensitive: understand where user is, and

give useful information and connectivity. If a user happens to be in Sydney looking for an ATM, the system as a whole should provide consistent information about the closest ATM machine, not for the ATM machine located in his/her hometown in London. At the same time, the system should be able to provide timely information about the FTSE index, when requested; the roaming and the use of local infrastructure to underpin “that” specific need of information and connectivity can achieve this capability (Ahonen, 2002; Benni, Laartz, & Hijartan, 2003; Beck, 2001; Andersson, Talborn, & Werkert, 2002; Deprez, Steil, & Dahlstrom, 2004; Neimeyer, Pak, & Ramaswamy, 2003; Tsalagatidou, 2000).

The following dimension is *moment*—more simply, time. It is therefore possible to manipulate passing time, planning, scheduling, and postponing or coping with a sudden need or

request, and even multitasking—that is, doing more than one thing at a time (talking on the phone while taking a picture of the monument in front of us to be sent to our friends at home, but two businessmen are passing by, talking about the next takeover of a company, so we have to catch up with our broker and tell him to buy more stocks of the same company).

The next dimension is *me*, a very personal area, but how much large it can be? Current mobile phones provide a quite remarkable amount of customization features (ring tones, colored screensavers, covers, and accessories to name a few): it is possible to think of more complex options, either serious or frivolous, and stretch the concept to an extension of our own personality. What is more, the content/service provided must be relevant for the user (example: do we really have to read through the whole horoscope or can we gain access to our own personal sign?). But this is not enough: the service as a whole should enable us to keep in touch with people we want to, give us the power to hold the keys (right now many mobile phones can ring different tunes according to the caller), or even allow many people to talk simultaneously. It easy to understand that *me* is the most powerful tool among the five (Ahonen, 2002; Benni et al., 2003; Beck, 2001; Andersson et al., 2002; Deprez et al., 2004; Neimeyer et al., 2003; Tsalagatidou, 2000).

Then we have *money*. This dimension is a further confirmation of the need/trend of converging technologies and converging multipurpose devices: a mobile phone that can easily replace wallet and coins, and enable mobile banking (many banks in several countries offer this service to their customers). On the business front, joining or offering this capability through sound alliances locally or globally can enrich the value chain. If this happens to be the chosen solution, it is more than ever necessary to adopt the tools described at the beginning of

this section because, from a practical point of view, there are either bottom-line issues (“how to share the pie” revenues) and KPI (key performance indicator) matters to define (and therefore the value proposition of each side).

The last dimension is *machine*: we have already seen on the market many mobile devices that can perform different tasks.

But What about the Marketing?

The previous section offers us insights that could have an impact on design, development, and execution of the marketing strategy. From the five lenses of mobile business, it is possible to define segments of users and classify them according to their behaviors and patterns of use—whether or not they receive and make calls; if they use local network or ‘roam’ because they travel a lot; in which day and time of the week; do they just speak or send messages, pictures, and data. Or whether or not they are business users or private users, how much is their bill, and how many times do they recharge their pay-as-you-go/prepaid mobile phones?

Table 8 sums up the possible combinations of segments on which to develop a marketing strategy; it is easy to understand that joining these patterns and the calculation of profitability (net contribution marketing) is more reliable (Ahonen, 2002; Benni et al., 2003; Beck, 2001; Andersson et al., 2002; Deprez et al., 2004; Neimeyer et al., 2003; Tsalagatidou, 2000).

Therefore the analyses is based, once again, on real behaviors. Let us consider, for example, a direct marketing campaign where your company has to deal with the mobile “stuff”. We can quickly design a top-down pyramid which includes all the cost/revenue situations, providing a sound metrics called “E:R” (expense to revenue) ratio for each media involved in this campaign (Roman & Hornstein, 2004).

Table 8. Marketing dimensions (for more details, see Ahonen, 2002)

OWNERSHIP	Personal phone, business phone, parents' phone, supplementary phone
CONTACTS	Few, many, random
LOCATION	National, overseas, movers, and shakers
TIME	Weekday, night, day, weekend
BEHAVIORS	Mostly caller, mostly receiver, sending data, images, files, music, call backer, avoider

The second consideration is that alliances and partnerships, as told before regarding the tools to use (Porter's 5 Forces and SWOT Analysis), enable a new definition of product/service in mobile business. Let us think, for instance, about Virgin, a global brand encompassing mobile services, travel, shops, music stores, to name a few fields filled by this company. What do we have at the end of this development process? A lifestyle proposition where customers and business cooperate to create a unique environment providing a unique experience: something which goes far beyond the so-called "+1 Factor" (a differentiating characteristic that makes customers choose one product instead of another) with a further positive impact on the bottom line, including brand equity (Ahonen, 2002; Benni et al., 2003; Beck, 2001; Andersson et al., 2002; Deprez et al., 2004; Neimeyer et al., 2003; Tsalagatidou, 2000).

The third consideration is that the overall value chain/value proposition needs to be properly reassessed, forcing all the parties to focus on what they do best: once again nothing new under the sun.

The last consideration is that, on the technical side, reusable technologies allow the roll out of quick and ongoing changes.

Current Trends

So, what is around the corner? Well, over the past few months, we have seen the number of mobile phone users overtake the number of

residential lines in many countries in Europe. At the same time, telecommunication companies are experiencing an increasing demand for high-speed fixed Internet (DSL, ADSL, broadband) for households. Moreover, massive advertising and promotion campaigns have been deployed to convince people to shift from the old traditional line to the modern and eye-catching fast lines for quite affordable prices (value for money approach), and bundled with new services and content. Basically we can consider this phenomenon as a migration of technology, business opportunity, and money into a more sophisticated environment: mobile phones are becoming a substitute for the fixed phone; the existing fixed line is changing to become the entry gate for more powerful capability; we are also witnessing the convergence between TV and computer in hardware and software. What it is more difficult to estimate is the churn rate among mobile phone operators: many of them cover residential, business, high-speed, and wireless worlds. Moreover, the price of mobile phones is decreasing (maybe they will be a commodity and the fight will be for the services and the applications enabled and provided to the customers). Lately, many refer to "walled garden" to explain the impossibility for a customer to use the services of another mobile provider unless he or she does not become customer. The fix solution is to subscribe to many contracts, have more than one mobile or at least more SIM card, and switch from one to another depending on the need of the moment. Once again the key suc-

Diagram 1. Value chain for the wireless business

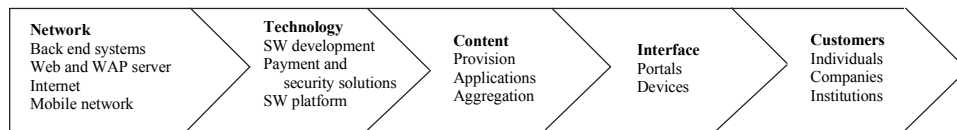


Table 9. Marketing dimensions revisited (Ahonen, 2002)

OWNERSHIP	Personal phone, business phone, parents' phone, supplementary phone
CONTACTS	Few contacts, many contacts, random
LOCATION	National, overseas, movers and shakers
TIME	Weekday, night, day, weekend
BEHAVIORS OF USERS	Mostly caller, mostly receiver, sending data, images, files, music, call backer, avoider

cess factor will be the quality of the service and its uniqueness, where customers will have strong bargaining power (Porter's 5 Forces are back again), and therefore the companies will need to find a strong value proposition to maintain their existing customers and attract new ones.

What Next?

Here we are at another crucial step: how to put the wheel on mobile business. Diagram 1 shows the value chain for the wireless business.

Now, let us try to swap the final step, "Customer," which consists of individuals, companies, and organizations, with Table 9.

The task is to develop a marketing-oriented strategy to match moody customers' demand. Why? Because happy customers return with frequent purchases; they are less available to change (the best way to keep a customer is to take him out of the market); they can be the most powerful marketing tool (word of mouth and referral have the highest impact among the different marketing and communication media).

Therefore there is an implicit high switching cost/barrier of entry in our favor—and finally, there is a tangible impact on the bottom line. It

is now possible to have a clearer understanding of the possible combinations of marketplaces, but this is not enough: we have already covered the importance of establishing strategic alliances and partnerships. Now we have to select, rank, and classify developing compatible metrics and KPIs for each step of the other parts. The next diagram should help to fill the gaps and have a clearer understanding. Think about your business and try to properly allocate your partner and suppliers by taking into account your customers. A humble set of clues to success:

- Do not think only about the technology issues, but consider the overall scenario (Plant, 1999), which is made by accounting (think about Dell: customers pay on the spot to buy a computer, but Dell pays suppliers 60 days or later, thus leveraging cash flow items), legal, distribution, and communication.
- Try to implement what Kalakota and Robinson (1999) suggest by teaming up or partnering with your competitors to reduce costs or gain a kind of advantage. (Are you surprised? You should not be, as

many other companies did the same such as Polaroid and the car makers including Covisint and others.)

- c. If you happen to read a marketing book, once you get to the chapter dealing with the “4 Ps,” remember that in our case, there are “7 Ps” (to product, place, price, and promotion, you need to include people, process, and physical presence).
- d. To gain a competitive advantage, think about your services/products and, once you have defined all their “ingredients,” do a displacement of one of them and combine those left accordingly (Kotler & Trias De Bes, 2003).
- e. Keep the cannibal in the family (Piercy, 2002)—be ready to cannibalize yourself before someone else does.
- f. Also be quick to reinvent your value chain. At the end of the game, you should have differentiated value propositions that can be further enhanced and improved.

A Thousand-Mile Trip Starts with a Single Step: But Where are We (Going) Now?

The potential for a mobile network is particularly high in the developing countries where it is difficult to set up a landline infrastructure for fixed Internet: for example, in the Philippines, SMS messaging has taken off rapidly.

Mobile devices are becoming mobile portals providing a wide range of customized services. The over-estimated 3G of mobile communication will—probably—deliver enriched experiences including sound and images and always be on (remember what we said about connectivity?).

On the business side, PDAs with diary, word processing, and e-mail features are becoming a killer communication tool; Mobiliss, an American-based wireless marketing company

has developed with Tribune Media, a wireless multiplayer game (by adapting Jumble, a scrambled word game) which can be an eye-catching advertising tool, as the words used to finish the puzzle can be redirected to advertisers’ products or services, and the customers can also access the advertiser’s call center (Harris & Dennis, 2002). Another example of advertising and promotion initiatives is given by PlanetHopper, a New York company which, after teaming up with General Cinemas theatres and a guide for bars and restaurants, is able to provide wireless coupons; now there are 20,000 users (Harris & Dennis, 2002).

If the sky seems to be pink, there are some clouds in the background: as we have already seen, there is the walled garden issue, the bandwidth, and the fact the mobile phones cannot compete with desktop computers in terms of display, computing power, and keyboard capabilities.

An interesting future development is SALT (Speech Application Language Tags), an initiative led by Microsoft, Intel, and Philips to realize a new platform enabling voice recognition-based services. It is still at an early stage but, on the business side, we can consider this project as a possible reengineering of the current call center. An English company, 365 Corporation, launched in 2001 the first comprehensive voice portal, Eckoh, which can be reached on the Web at <http://www.08701101010.com> or by phone (mobile or fixed). With Eckoh, users are able to e-mail, shop, listen to news, arrange conference calls, set up appointments, and more.

Some Tips About the Current Business Environment

An old Latin phrase says “Dum Romae consulitur, Saguntum expugnatur,” (Livy.): translation “while in Rome the Senate is dis-

cussing, the enemies are conquering Sagunto” (a city in Spain). You may pose this sentence any way you like, but the meaning is quite straightforward: What is going on at the shop level NOW? What does the situation look like?

The Consumers’ Behavior: Unpredictable Bills Make Them Move Away

A recent survey carried out by McKinsey consultants (Mc Kinsey Quarterly, 2004) tells us that “Mobile-telephony subscribers give many reasons for switching operators—prices, brands, friends’ recommendations—but very few cite highly volatile monthly bills.” But “subscribers whose monthly bills fluctuate substantially tend to churn much more than people with more consistent bills.”

With an acid test on these movers and shakers’ customers, operators could reduce the churn rate by “introducing pricing plans solely for subscribers with volatile bills”: they could subscribe to an annual contract which includes a monthly predefined rate based on their previous pattern/phone bill, receiving timely alert as to any remarkable overdraft, thus enabling them to balance their position at the end of the contract.

This strategy should be combined with a more comprehensive program to “rise switching costs” from one operator to another (currently if you change operator, you can take the number but not, for example, the credit of your pre-paid mobile phone—another example of “walled gardens” approach).

Customers’ Pain: Why Do They Complain?

The honeymoon between customers and telecommunication companies is over: during the past few years in Europe, the telecommunica-

tion sectors have been under the control of governments. The road to privatization and competition has been opened in the U.S. by a massive antitrust law case that gave birth to the so-called “baby bells”; Europeans followed slowly, and even though the technology gap between the Atlantic rims is almost filled, particularly in the last few years, consumers/subscribers are still facing a troubled situation with unclear terms and contracts, inaccurate bills, and unfair behaviors. Moreover, the dawn of new and converging technologies—where the same company is or can be provider of landlines, mobile, broadband, and interactive TV either as technology or content provider—makes the situation even more complicated.

But Why Do Customers Complain?

A complete survey was carried out late in 2004 by the Better Business Bureau (BBB) in the U.S. The majority of incidents involved billing problems, structured into three major categories:

1. setup and access, including difficulties retrieving statements online or getting detailed call logs;
2. errors, many involving calls made while phones (according to their owners) were not in use; and
3. failure of statements to reflect terms, such as credit or rate-plan changes, negotiated in discussions with customer service agents.

This has very little to do with technology, but with the questionable strategy carried out by many companies to outsource or employ contractors in customer-faced roles: in many cases these employees are not adequately trained, motivated (and paid of course) to do a good job and provide a good service.

Another set of complaints involves cases in which the customer believed that the carrier acted deceptively or otherwise misrepresented the terms of a contract. Such incidents are referred to as “miscommunication” to avoid the question of whether promises had really been made and broken or the customer misunderstood terms that were actually well explained—for instance, when a store’s sales agent tells a customer that a certain house lies within a carrier’s coverage area, without explaining that the quality of coverage varies within that area. This scenario is confirmed by the McKinsey survey (McKinsey Quarterly, 2004):

All of these issues are aggravated by the physical separation between the channel where the customer relationship originates (generally a store, and not always one under the carrier’s direct control) and the channel (usually a call center) that provides customer service for the carrier. The different objectives of the managers who run these disparate operations can be tricky for the carrier to reconcile.

So, what to do? Customer service systems (called CRM, if you like) must provide agents with a customer’s entire history of contacts from all channels. Agents should promptly refer calls to their managers when requested to do so, and managers must provide coaching for them in order to prevent rudeness. Most important, at the outset of every relationship, the carrier should communicate clearly to the customer how the contract period works. If it later changes, the carrier should explain the impact promptly.

Fixes require a mobile-telecom company to coordinate its efforts across departments: customer service, IT, marketing, and retailing (carefully select your partners at the shops level if you are involved in indirect sales). You will

need to improve communication among these functions and to devise appropriate incentives for each of them (and, if any, replace the bad performer as soon as possible). And there must be clear consequences for stores that repeatedly fail to give new customers full explanations of the carrier’s policies and pricing plans (once you select stores and more generally partner to complete your value chain, establish clear and well-understood KPIs in a carrot-and-stick way). Given the tremendous benefits of heading off intense customer dissatisfaction, companies should recognize the high cost to tackle these problems. In other words, preventing a problem is the best defense.

No-Frills and Low-Cost Operators

Denmark was among the first places to experience the trend. In 2000, Telmore bought unused mobile capacity from incumbent TDC Mobile. Telmore targeted college students with a fixed-price offer providing voice and Short Message Service (SMS) at rates that were initially more than 20% below those of the competition. Telmore minimized costs by using the Internet as a distribution channel, backed by a small call center, which resolved questions for a fee. This stripped-down approach eschewed expensive product offerings, mass advertising, and subsidized handsets. Despite the bare-bones offer, Telmore was consistently rated highest in customer satisfaction among Denmark’s mobile providers. With an initial investment of just a few million euros, it captured almost 20% of the consumer part of the national mobile market (Beck, 2001; Braaf, Passmore, & Simpson, 2003; Deprez et al., 2004).

Germany is Europe’s largest mobile phone market with 60 million subscribers, a fact that illustrates the broader interest in the no-frills option. A survey of thousands of German mobile users was conducted in order to estimate

the importance they attach to the services and applications that wireless companies typically offer. The outcome shows that a third of the market has limited interest in advanced features or personal interaction with mobile operators: these customers want to make simple phone calls (and perhaps to use SMS) and would be happy to stick with handsets they already own or to pay for new ones. Since traditional mobile operators reduce up to 200 Euros a customer in handset subsidies, this finding is particularly important: no-frills customers do not want the large variety of products and services, such as access to news and weather bulletins, that are included by default in some standard contracts (Beck, 2001; Braaf et al., 2003; Deprez et al., 2004).

Attackers in several countries are experimenting with variations on the no-frills theme. Some of these players, such as Comviq (Sweden) and Telering (Austria), own their infrastructure, while others like CBB Mobil (Denmark) and Telenor's djuice (Sweden) buy capacity from other operators. Aggressive competitors have also made Hong Kong a price-sensitive market, and the U.S. is seeing some initial paths. In emerging markets like Asia and Eastern European countries, customers consider mobile phones as a substitute for fixed-line services, so the mobile market has matured rapidly, increasing the growth of price-sensitive segments; low personal incomes and the operators' need to pare costs to the absolute minimum have intensified the trend, and do not forget the digital divide issue for the developing countries and how mobile communication can fill the gap with developed countries (Beck, 2001; Braaf et al., 2003; Deprez et al., 2004).

Which are the drivers for this phenomenon? Several factors:

- household income,
- exposure to discounters, and
- Internet penetration.

In addition, advertising that focuses on price generally increases a market's sensitivity to it. The ability to use the same handset with different providers also makes markets more conducive to no-frills plans. Saturated markets with significant overcapacity tend to embrace them, since attackers can buy low-cost unused capacity from smaller players (Beck, 2001; Braaf et al., 2003; Deprez et al., 2004).

What happens next? As long as mobile phone markets mature and become saturated, the proportion of customers unwilling to pay for anything beyond basic services will keep growing. Incumbents must balance the potential threat of the attackers and decide whether and how to deal with the needs of these growing segments. The incumbents must improve the effectiveness and efficiency (the "do the right thing and do it right" motto) to stay cost competitive while facing declining prices. Many operators will invest in customer lifetime-management systems. Eventually larger operators will adopt more sophisticated market segmentation approaches (Beck, 2001; Braaf et al., 2003; Deprez et al., 2004).

CONCLUSION

Over the previous pages, we tried to outline the opportunities and the challenges of mobile business by describing and analyzing some of the most interesting initiatives, providing at the same time a possible set of tools to use to design and develop wireless solutions with a spot on the business aspects and challenges, perhaps the most challenging and intriguing part.

A recent survey (Kotler, 2004) highlights that anyone of us is hit by approximately 1,500 advertising messages everyday, and in this race the winner is the one able to provide fast, customized, and reliable solution. The DHL example, some pages ago, confirms that a mobile implementation can be a success: later

in this book you will also find a collection of companies who successfully developed a mobile strategy—additional evidence that “wireless” does not mean “for the big guys only.”

Now it is your turn: stop reading and try to do something.

REFERENCES

- 365 Corporation. (2004). Retrieved August 26, 2004, from <http://www.08701101010.com>
- Ahonen, T. (2002). *M-profit. Making money from 3G services*. New York: John Wiley & Sons.
- Ahonen, T. (2003). *Services for UMTS: Creating killer applications in 3G*. New York: John Wiley & Sons.
- Andersson, T., Talborn, H., & Werkert, M. (2002). *Business models for mobile Internet*. Lund, Sweden: Institute of Economic Research, Lund University.
- Bayne, K. (2002). *Marketing without wires*. New York: John Wiley & Sons.
- Beck, H. (2001). Making money where it's scarce. *McKinsey Quarterly Special Edition on Emerging Markets*. Retrieved from <http://www.mckinsey.com>
- Benni, E., Laartz, J., & Hijartan, K. (2003). The IT factor in mobile services. *McKinsey Quarterly*, (3). Retrieved from <http://www.mckinsey.com>
- Better Business Bureau. (2004). *BBB advice: Don't ignore bogus billing during holiday rush*. Retrieved October 14, 2005, from http://www.frostillustrated.com/news/2004/1229/Consumer_News/014.html
- Braaf, A., Passmore, W. J., & Simpson, M. (2003). Going the distance with telecom customers. *McKinsey Quarterly*, (4). Retrieved from http://www.mckinseyquarterly.com/article_abstract.aspx?ar=1356&L2=22
- Bughin, J. R. (2004). Using mobile phones to boost TV ratings. *McKinsey Quarterly*, (4).
- Deprez, F., Steil, O., & Dahlstrom, P. (2004). Meeting the no frills mobile challenge. *McKinsey Quarterly*. Retrieved October 15, 2004, from <http://www.bangkokpost.com/mckinsey/McKinsey211004.html>
- Euresco. (2002). *Mobile electronic commerce*. Emporio—Project P1102—European Institute for Research and Strategic Studies in Telecommunication, Germany.
- Eylert, B. (UMTS Forum Chairman). (2001, September 25). 3G chances and market opportunities. *Proceedings of the UMTS Forum*, Singapore GSM-GPRS in the Asia Pacific.
- Gratton, S. J., & Gratton, D. (2004). *Marketing wireless products*. London: Elsevier Butterworth Heinemann.
- Harris, L., & Dennis, C. (2002). *Marketing the e-business*. London: Routledge E-Business.
- Hoque, F. (2002). *The alignment effect*. London: Financial Times Publisher.
- IBM Institute for Business Value. (2004). Mobile portal strategy: When did business partnership become so critical to customer value? White Paper, IBM, USA. Retrieved August 26, 2004, from http://www-8.ibm.com/services/pdf/IBM_Consulting_Mobile_portal_strategy_Is_collaboration_the_key_to_customer_value.pdf
- Kalakota, R., & Robinson, M. (1999). *E-business 2 roadmap to success*. Reading, MA: Addison-Wesley.
- Kotler, P. (2001). *Marketing moves*. Englewood Cliffs, NJ: Prentice-Hall.
- Kotler, P., & F. Trias De Bes. (2003). *Lateral marketing: New techniques for finding*

- breakthrough ideas*. New York: John Wiley & Sons.
- Luftman, J. (2002). Appendix F. Hoque (Ed.), *The alignment effect*. London: Financial Times Publisher.
- May, P. (2002). *Mobile commerce*. London: Cambridge Press.
- Neimeyer, A., Pak, M., & Ramaswamy, S. E. (2003). Smart tags for your supply chain. *McKinsey Quarterly*, 4.
- Newell, F. (2004). *Why CRM doesn't work*. London: Hogan Page.
- Paavilainen, J. (2002). *Mobile business strategy*. Reading, MA: Addison-Wesley.
- Panis, S., Morphis, N., Felt, E., Reufenheuser, B., Boem, A., Nitz, J., & Saarlo, P. (2002). *Mobile commerce service scenarios and related business models*. Eurescom Project P 1102, Emporio—Project P1102—European Institute for Research and Strategic Studies in Telecommunication, Germany.
- Piercy, N. (2002). *Market-led strategy*. London: B&H.
- Pippow, I., Eifert, D., & Strüker, J. (2002). *Economic implication of mobile commerce—An exploratory assessment of information seeking behavior*. White Paper, University of Freiburg, Germany.
- Porter, M. (1980). *On competition*. Boston: Harvard Business Review Books.
- Roman, E., & Hornstein, S. (2004). *Opt-in marketing*. New York: McGraw-Hill.
- Siemens Application Marketing. (2002). *Mobile business—A task-based focus*. Retrieved from http://www.verista.com/mbusiness/mbusiness_wp_001.pdf
- Tsalagatidou, A., & Veijalainen, J. (2000). *Mobile commerce emerging issues*. White Paper, Department of Computer Science and Information Systems/Information Technology Research Institute, University of Jyväskylä, Finland.
- White, J. (2001). *Enabling e-business*. New York: John Wiley & Sons.

Chapter XXXIII

Relating Mobile Computing to Mobile Commerce

Nina Godbole

*CQA, CISA, PMP, CSTE, ITIL (Foundation) Certified Professional Member—
Computer Society of India, India*

ABSTRACT

In today's digital economy and the extended enterprise paradigm, mobility is on the rise. It is important to perceive mobility as an opportunity, rather than a threat. Although m-commerce is still at its infancy, it serves as an extension to e-commerce sites—it has been regarded as a value-added service. However, there are many issues and challenges while reaping full benefits of mobile computing solutions for m-commerce. This is because mobility and mobile computing is replete with many challenges on the business front, technical challenges as well as social challenges. This chapter undertakes discussion on understanding mobility and categories of mobile user types, understanding the meaning of m-commerce. Typical applications that support the m-commerce paradigm are illustrated through case studies. The chapter ends with a discussion on legal implications of mobile technology, and future directions for mobile commerce and mobile computing. The key message is that mobility is not just about connectivity—it is about function it provides and the way organizations work in today's digital economy.

INTRODUCING MOBILE COMMERCE

The Internet, especially combined with wireless technologies, has become more than just a communication media. Together, they form important business drivers for e-commerce (electronic commerce) and m-commerce (mobile commerce) and, as such, have become

integral features of the global economy. The Internet, combined with wireless communications, enables flow of information among business players, increasing the speed and accuracy with which businesses exchange information and, simultaneously reducing the costs of transactions. For example, Internet economics facilitate a reduction in the number of middlemen involved. Along with the development of

the Internet is the phenomenal growth of the mobile communications industry. In fact, the mobile communications industry is growing so rapidly that in 1999, there were more mobile phones sold than automobiles and PCs combined (Telecommunications Service Inquiry, 2003). During the Qualcomm BREW 2005 Conference (statistics on camera phone), a phenomenal rise in the use of SMS and picture messaging was reported. Kerner (2005) also states that by 2009, consumers worldwide will be buying and/or replacing their mobile phones at a rate of one billion per year. Thus, it should be anticipated that the trend in e-commerce, which has influenced the economy and lifestyle of today's culture, will extend to m-commerce enabled by the wireless technology and the wireless application protocol (WAP).

According to a Commonwealth Report (Telecommunications Service Inquiry, 2003), it is expected that by the year 2006, there will be 923 million Internet users, whereas 543 million will be mobile (wireless) users. Currently, there are almost 8.5 million mobile services in Australia after 13 years of operation, compared with approximately 10.64 million fixed lines after over 100 years of operation. The number of mobile phones sold in Europe has also grown rapidly. In the UK, there are approximately 30 million subscribers. In both Italy and Finland, 70% of the population owns a mobile phone. Consequently, this global trend has motivated businesses to adjust them with the change. Gartner raised its 2005 forecast (see Kerner, 2005) for mobile phone sales to 779 million units, an increase of 16% over 2004. As recently as May 2005, Gartner revised its 2005 mobile phone sales forecasts from 720 million to 750 million units, which is a 13% increase over 2004.

Thus, the mandate to mobilize business data is clear and, with global m-commerce reaching \$200 billion by 2004 (Gartner, 2005), m-commerce is the new benchmark. Just as business

Figure 1.



organizations compete with one another to do business on the Internet, the evolution of technology continues through handhelds and smart phones. Examples of such handheld gadgets are shown in Figure 1.

The benefits of *pervasive information access* and the ability to do *business anywhere* are evident in organizations that have decided to embrace mobile commerce opportunities as part of their strategic choice. This poses certain critical issues for today's CEOs and CIOs: How will my company measure up against the m-commerce benchmark? Are we supplying business data to our workers via handheld devices? Are we supporting our critical business functions with mobile technology? Is our IT shop providing support to end users? These are some of the crucial questions explored in this chapter. This chapter undertakes discussion on understanding mobility and categories of mobile user types, understanding the meaning of m-commerce, and typical applications that support the m-commerce paradigm. Finally, the chapter ends with a discussion on legal implications of mobile technology, and

future directions for mobile commerce and mobile computing.

THE NEW M-COMMERCE SCENARIO: THE MEANING OF MOBILITY

With the advances in the Internet, organizations make greater attempts to extend their “virtual reach.” This results in greater mobility of the workforce. Thus, there is the need to deal with the requirements of employees as they move around (different locations in the building, between buildings, between companies, and while in transit traveling from one country to another). In the “Extended Enterprise” scenario of today, this provides immediate benefit in terms of workflow efficiencies. However, this also introduces problems with how we manage access and information sharing with suppliers and other external resources. “Mobility” is a means to change the way we do business. Thus, *mobility* is an attempt to extend corporate networks to connect to people who are increasingly on the move.

Mobility: The New Challenge

Over a period of time, “mobility” has affected us in a slow but sure way. One can argue that the workforce has always been mobile—commuting to and from work, completing sales transactions at a customer’s site using handheld devices to access business information held on back-end databases at their parent organization, meeting suppliers and prospects, and making instantaneous changes to negotiating stances aided by access to customer history information accessed through mobile computing devices while working in the field. After all, organizations do have groups of geographically dispersed skills. However, today’s workforce

does more than this in the paradigm of m-commerce aided by mobile computing.

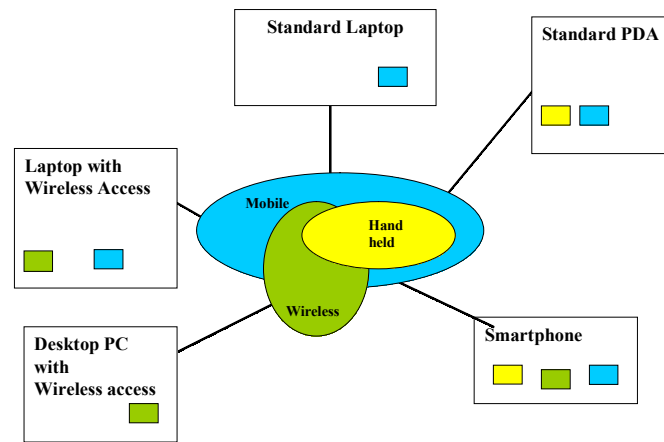
Thus, today *mobility* impacts the workforce in a major way, as we have become more dependent on technology, the need to remain in contact, the need to be able to provide ever more up-to-date data and to be able to beat the competitors in the fast-moving marketplace.

Earlier in this chapter there was a discussion on the meaning of “mobility.” Mobility can be considered as an opportunity for organizations to outpace the competitors, to add a differentiation by bringing a change to the style of working, to utilize the resources at our disposal in an optimum manner, without worrying about the underlying issues of technology. A June 2003 report commissioned by GRIC posed some interesting questions to organizations as part of a survey. A summary of the report is as follows:

- **Budget Spend:** The question posed in the survey was, “What is the percentage of overall mobile computing budget spent by on mobile applications and services?” A large number of respondents said that this was between 31% to 50%.
- **Priority Perceptions:** Another question posed in the survey was regarding “perceptions on mobile commerce and mobile computing. From the responses (in terms of priority for remote wireless access to various applications), it emerged that priorities in terms of strongest business case appear to be towards applications that cover *mobile office, sales force automation, automation of professional services, field service support, logistics, and transportation*.

When it came to the survey question on “*the relevance of third-party mobile applications and services to organizations,*” *m-commerce* and *multimedia messaging* were found to be

Figure 2.



on top, while *collaboration* and *location-based computing* seemed to be rated much lower.

WHAT IS MOBILE COMMERCE?

Let us now try and understand mobile commerce in greater detail. M-commerce refers to the purchase of products and services using a mobile terminal. With the success of e-commerce for many types of purchases, m-commerce is considered the next logical wave after e-commerce supporting “business on the move.”

Although it may appear that essentially m-commerce is e-commerce without the constraints of wired connection, many believe that m-commerce introduces time-location independence that never existed before—that is, the freedom to compute *anywhere, any time*. The “context sensitivity” feature is believed to be another added advantage to mobile commerce in the mobile computing paradigm. For example, according to Kolari et al. (2005), when mobile services and applications are made ‘context-aware,’ they offer contextually relevant information to the users.

It is important to discuss *handheld devices* in the context of *wireless computing*, which is one of the main enablers of m-commerce. This helps us provide clear definitions for mobile computing and wireless computing. Figures 2 and 3 help us understand that.

Mobile commerce is best suited where the consumer is driven by a “sense of urgency”—that is, when they need to have their goods and services immediately for upcoming functions and events. Of course, logistic issues are the constraining factors in an m-commerce scenario, in the sense that having completed a purchase transaction using his mobile/wireless device, the consumer still has to wait for the physical delivery of the product. However, for certain kinds of products purchased through m-commerce, this limitation is overcome, for example, a movie ticket or information services, e-brochures, and so forth.

MOBILE COMMERCE ENABLERS

The previous section discussed what mobile computing is all about. This section addresses what promotes mobile commerce. Mallick (2003)

writes of five major driving forces that act as enablers for the adoption of mobile and wireless solutions. In this section, we offer a brief overview of these factors:

- Wireless Networks
- Mobile Devices
- Software Infra-Structure
- Standardization
- Mobile Internet

Of the factors indicated above, the last one deserves a longer discussion. It is a “mega-enabler” for m-commerce in the mobile computing paradigm.

Wireless Networks

The initial marketing focus for *wireless carriers* was aimed at the consumer market. The emphasis was on inexpensive handsets and affordable calling plans for consumers. However, this conceived market did not generate as much revenue as expected, and since then, much of the focus of the wireless data services has moved to the corporate market. Currently, some significant improvements are in the offing for making wireless network improvements. They are summarized below. All these factors offer a great boost to mobile commerce.

Increased Bandwidth

Next-generation wireless networks will overcome the data capability limitations of voice-oriented wireless networks. These new generation wireless networks are designed for wireless data, and provide communication speeds between 56 Kbps and 384 Kbps. This speed is very adequate for the limited amount of data that is typically required from a wireless device.

Always-On Capability

This term refers to users’ ability to access data at any time, without having to establish a connection to the wireless network for each session. This ability has been enabled by packet-switched networks. Users welcome this because, whereas for the *circuit-switched networks*, billing is on the basis of amount of time they are connected, *packet-switched networks* charge the users based on the amount of data transferred.

Lower Costs

This supports the above point and results from packet-switched networks allowing wireless operators to provide new offerings that are based on data usage instead of call times. Another factor is roaming charges. Many carriers have drastically reduced roaming charges for using other carriers’ services.

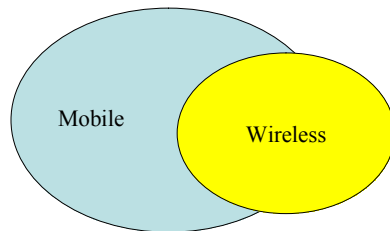
Enhanced Services

New services are being offered that can add value to mobile solutions. Many carriers now allow users to download additional applications for their devices. In addition, some integrated services, such as *location-based services*, are providing companies with valuable features that can further contribute to the success of mobile solutions.

Inter-Operability between Carriers

Wireless operators are starting to work in a collaborative fashion to help promote benefits of mobility. The result is a new level of inter-operability for both data and voice communication. An example is the ability for users in North America to send text messages to users on

Figure 3.



In most cases, “Wireless” is a subset of “Mobile”, but in many Cases, an application can be mobile without being wireless

other networks, a feature made available in 2002.

Mobile Devices

Incredible advances are being made for mobile devices. A few years ago, the choice was between a wireless phone and a simple PDA. Now there is a long list of options ranging from high-end PDAs with integrated wireless modems down to small phones with wireless Web browsing capabilities. Even the simplest of handheld devices provides enough computing power to run small applications, play games, and of course make voice calls.

A key driver for the growth of mobile solutions is the proliferation of devices in the enterprise. As more personal devices find their way into the enterprise, corporations are realizing the benefits that can be achieved with mobile solutions. The trend is for smaller devices and more processing power. Some such devices are shown in Figure 4.

Most of the devices depicted in Figure 4 work like “pocket PC,” and they are very powerful. Users can count on them when it comes to portability, data security, and maximum operating time—crucial features for devices used for mobile commerce. Most of the devices shown above are featured by high-performance processors (with speeds ranging

Figure 4.



up to 200 MHz), color display with front light, at least 64 MB of RAM, and 64 MB of nonvolatile Flash storage to protect data from resets or even complete power loss. Thus, today a device that fits in your hand (i.e., “handheld devices”) has as much computing power as desktops did 10 years ago. The same device is also able to communicate over a wireless network and view office documents at the same time. This combination of size, power, and flexibility is a key enabling technology for enterprise mobile solutions. In today’s mobile commerce paradigm, when employees are empowered by mobile solutions, a number of benefits emerge: increased employee productivity, faster response times to business changes, streamlined business processes, improved customer satisfaction, increased competitive advantage. A section on case studies later in this chapter illustrates this.

Software Infrastructure

To make a mobile computing solution successful, a strong software platform is required along with wireless networks and mobile devices. These platforms offer support for leading mobile computing models such as wireless Internet,

smart client, and mobile messaging architectures. Software vendors have made significant advances in several key issues that enable mobile application development. For wireless Internet applications, the introduction of wireless application server frameworks, device emulators, WAP (wireless application protocol), and development tools has allowed many m-business solutions to be possible.

Standardization

As per industry reports (Quocirca, 2004), many organizations have cited the lack of standards as a major obstacle in developing mobile solutions. Organizations worry that the technology they use will become obsolete before their mobile computing applications get deployed.

This is a valid concern because use of established standards provides an assurance that mobile computing applications will work now and in the future. However, as with any new industries, it takes time for standards to develop, especially in a continuously evolving scenario such as mobile computing for m-commerce. We are now at the point where standards are emerging at all levels of the mobile environment.

Mobile Internet

This is a phenomenon to reckon with and hence worth discussing in this section. There are many disputes and debates around this topic. For example, take iMode. It is an innovative solution that has revolutionized the Mobile Internet. It focuses on enhancing customers' lifestyles by providing users with an intuitive user interface for easy access to a wide variety of lifestyle Internet services over dedicated iMode-enabled handsets. These include multimedia content, push e-mail services, games,

mobile commerce, and banking. All this is integrated on an iMode gateway that will be developed especially to allow iMode users to have an optimized, end-to-end customer experience. Essentially, iMode, like WAP, is a way of providing information to mobile devices. iMode, however, is slightly different from WAP in that it uses cHTML—that is, compact HTML (<http://www.webopedia.com/TERM/C/cHTML.html>)—as a markup language, and uses more traditional Internet protocols to deliver it. The content is served using HTTP to a so-called iMode center, which in a way works like the WAP gateway. The iMode resource center (<http://www.palowireless.com/imode/abouthtml.asp>) states that unlike WML used for WAP devices, c-HTML is a subset of HTML that leaves out coding for JPEG images, tables, image maps, multiple character fonts and styles, background colors or images, frames, and cascading style sheets. These things are excluded due to the low bandwidth and limited screen size of cell phones. Not only is cHTML simpler than WML for WAP phones, but developers need only make one version of the site for all iMode devices.

Thus, iMode brings easy-to-use mobile data services to the user's fingertips in a transparent manner. For example, users need not be technology experts to use iMode services to make ticket reservations or explore mobile banking, stock trading, or restaurant bookings on the mobile phone. iMode was introduced in Japan in 1999; today it is the world's leading mobile Internet service. According to one source, (<http://jgohil.typepad.com/mineternet360/imode/>), iMode currently offers 6,700 official sites worldwide with thousands of applications for the enjoyment of over 46 million mobile users worldwide. In Japan, in addition to the official sites, there are more than 80,000 independent iMode sites. iMode is a household name in Japan and the phenomenal success of iMode there pro-

vides a tested and proven ecosystem comprising close partnerships and strong associations with content providers, handset manufacturers, backend support partners, and global iMode alliance operators. The iMode platform enables content providers to develop or adapt their Internet content and applications easily for iMode devices.

The iMode service is provided on an open platform based on de facto Internet standards. To sum up, its easy-to-use functionality has attracted countless content and application providers, stimulating a vibrant market with thousands of sites to meet the needs of customers' diversified mobile lifestyles.

Although the iMode technology breakthrough may look promising, there are a number of problems. For example, consider the following from CapGemini (2001):

- In Europe, nearly 26% of employees in large companies have a GSM (Global System for Mobile Communication) handset for professional use, which is now higher than the rate regarding laptop equipment (16% of employees) and much more than equipment in personal digital assistants (PDAs) (4%).
- Voice usage is still dominant: more than two-thirds of companies have at least one WAP handset as a privilege for high-level managers and, on average, only one employee out of 24 has such a device. Furthermore, less than half of the companies equipped with WAP use it for data applications.
- WAP is used mainly for basic applications such as messaging, schedule consultation, or intranet access. Vertical applications such as maintenance, CRM, ERP, fleet management, and so on have been implemented only in a few companies.

- About one-third of the companies have already implemented Mobile Internet applications; 50% are said to be doing this in the near future.

The main reason for implementing Mobile Internet solutions is to facilitate exchange, notably through messaging. Indeed, many implemented applications are based on SMS solutions. But Mobile Internet is also used by some pioneer companies for advanced functions such as CRM, e-procurement, marketplaces, or supply chain. Even if only a few have implemented such applications, most companies are investigating Mobile Internet for the above mentioned functions in order to increase productivity and to acquire new clients.

The integration of Mobile Internet applications in existing information systems and middleware requires substantial efforts from companies and suppliers. As for the expectations, concerns of the users tend to be mainly around the types of devices, debit rates, tariffs, costs, and security. This is summarized below:

- In terms of devices, laptop computers are considered as the most suitable terminals, followed by cellular handsets and PDAs. Organizations may not want a unique device. It depends on the type of application and the work locations of the employees.
- Current data rate is not sufficient, but there is no immediate need for high speed; 60 kbit/s could be enough for most users, and consequently, GPRS (general packet radio service) could satisfy a large part of the market in the short term.
- For tariffing, basic voice services should be free of charge and companies would prefer a billing by volume of data or flat rate.
- In terms of investments, development costs are a major concern for most companies.

The cost of terminals is also seen as an obstacle for the smallest companies.

- Finally, security is a big challenge, as users want to be sure that all elements are protected (device, network, gateway, and server).

In the United States, the situation is quite different as the base of cellular users is significantly lower than in Europe (only 40 subscribers per 100 inhabitants), and the approach towards Wireless Internet is very pragmatic. Business users look for comfortable, easy-to-use, and relatively cheap solutions. The development of wireless packet network connections to PDAs is illustrative of that. As for operators, they prefer to upgrade their existing networks in the short term. Kerner (2005) from Gartner is also reporting that the PDA shipments are on the rise.

In Asia, there is a potential good outlook for Wireless Internet, with Japan and South Korea leading the market. In Japan, there are now more than 37 million Wireless Internet users (22 million users for NTT DoCoMo's iMode). However, in most cases there is no clear differentiation between residential and business users in terms of marketing approaches and offers. In the medium and long term, the Chinese market should create a strong dynamic as the cellular base is increasing rapidly (and should serve as a basis for advanced services, rather than fixed network).

Thus, to summarize, on the Mobile Internet, the key issues and challenges are (Quocirca, 2004):

- The success of Mobile Internet will be decided on three frontiers—technology, devices, and design quality of mobile commerce applications.
- Major players such as Ericsson are playing a key role in the mobile application

initiative as part of their strategy. Major stakeholders of the Mobile Internet community are: technology corporations, device/terminal manufacturers, software developers, mobile operators, content owners, ISPs and ASPs, start-ups, and venture capitalists.

M-COMMERCE APPLICATIONS IN MOBILE COMPUTING SCENARIO

Having discussed mobile commerce enablers, the power of mobile devices, and important phenomena such as the Mobile Internet and so forth, this section looks at the typical mobile computing solutions that drive the m-commerce market, including:

- **Digital Purchases:** These are products that can be downloaded and used immediately. Some of these m-commerce transactions are already happening (e.g., games and ring tones for cellular phones). The advances in mobile handheld devices make them great vehicles for game playing.
- **Mobile Banking:** Wireless devices provide two benefits to mobile banking. The first is providing access to personal bank accounts to view account history and execute transactions. This is an extension to Internet banking that is already very successful. The second benefit comes from use of mobile devices for online payments through e-cash.
- **Information Services on the E-Tap:** In today's dynamic world, "fixed-locations" are losing their meaning. Mobile users often feel out of touch with their daily routine as they travel around the globe in pursuit of business activities. Information services help address this need by providing information that the user is accus-

tomed to having. Typically, this information is stock quotes, weather information, entertainment, sports scores, and so forth. With mobile messaging technology, many forms of information can actually be pushed to the user in the form of an alert or notification.

- **Location-Based Services:** These offer the ability of the merchants to capture and provide services based on a user's current location and requirements are a new and powerful approach to selling services. Location-based services allow consumers to find the precise information they need at the exact time they want to use it. This will be an important enabler for m-commerce solutions, although privacy concerns will have to be addressed before location services enjoy wide acceptance in the user and business community.
- **Mobile Shopping:** Most forms of shopping are not going to be popular soon from the mobile devices. The reasons for people's discomfort using mobile devices for shopping is mainly due to the plethora of technical problems faced by mobile devices. For example, it is impractical to surf for items using constrained devices, while other methods of shopping are much more productive and socially enjoyable. Interestingly, there are some forms of purchases that lend themselves well to m-commerce. For example, purchasing travel and movie tickets for the same day. This form of m-commerce is found very convenient by users. Mobile devices can also be used for "comparison shopping" before making a purchase: a shopper in a retail store may want to first find the current price of a product from an Internet vendor to ensure they are getting a good price.
- **Mobile Advertising:** As mobile users start taking advantage of m-commerce solutions, mobile advertising is sure to follow. The mobile operators have access to several types of information that is attractive to advertisers. Once the advertisers find out where the users are located and what they use their mobile phones for, advertisers can send out personalized messages to these prospective customers. There are a number of threats for m-commerce performed on mobile handheld devices. Most notable among them is the risk of "customer backlash." This is the biggest obstacle to mobile advertising. If users start getting unsolicited messages (equivalent to "spam mails") and advertisements on their handheld/mobile devices, they are likely to switch service providers or keep their mobile devices off most of the times. Annoyed users may even go to the extent of stopping the usage of their handheld device. For this reason, in the near future, there could only be requested advertisements, such as the nearest gas station or restaurant.

Mobile User Types

Various sections, so far, discussed the meaning of mobility in the new m-commerce paradigm. The concept of mobility has a history. Mobility does not cover only telephony wireless standards, such as GSM and GPRS. Other technologies, too, have emerged. This includes the high-speed wireless networks of WiFi/802.11 and the Personal Area Network technology, Bluetooth. For the truly mobile user, these wireless technologies also need to be supported by "tethered" technologies including telephony and network usage, for example in hotels, in

transit areas (e.g., airport lounges) and in customer/supplier offices.

In view of this background, mobility needs to be considered outside the constraining realms of the underlying technology. It is important to understand that mobility *is a way of working, not just a way of utilizing available technology*. Amongst others areas, mobility must cover those working from home, on the road, in hotels, and at airports, as well as those moving around within the organization boundaries. A mobile solution must provide not only the technology to provide access—it must also provide the functionality for the user to add value to the company.

Under this paradigm, “mobile solution” can be defined as:

The meeting of an individual's or a group of users' functional needs, enabling a company to gain added value through optimal utilization of each individual's time while they are not at a normal fixed work desk.

Within this definition, four main types of mobile users can be defined:

- **Tethered/Remote Worker:** This is considered to be an employee who generally remains at a single point of work, but is remote to the central company systems. This includes home-workers, tele-cottagers, and in some cases, branch workers.
- **Roaming User:** This is either an employee who works in an environment (e.g., warehousing, shop floor) or in multiple areas (e.g., meeting rooms).
- **Nomad:** This category covers employees requiring solutions in hotel rooms and other semi-tethered environments where modem use is still prevalent, along with the increasing use of multiple wireless technologies and devices.

- **Road Warrior:** This is the ultimate mobile user—spends little time in the office, but requires regular access to data and collaborative functionality while on the move, in transit, or in hotels. This type includes the sales and field forces.

For these new actors of *extended enterprise* in the world of mobility, the tools that are required to create an effective mobile infrastructure revolve around the need for communication and collaboration. Within this space, we need to consider the likes of e-mail, calendaring/scheduling, contact management, discussion forums, virtualized meetings, and so on. While the majority of these systems can be provided in-house, we should ask ourselves whether we should be providing them in this way. If we are looking to control costs and provide an optimal solution, then we need to consider opportunities for minimizing the skills required to manage these technologies, and removing our dependence on the need for application patches, upgrades, down-time, and so on. Given the emerging emphasis on “focusing on the core competency,” today's organizations are looking for outsourcing opportunities to third parties for management of mobile computing/m-commerce solutions.

For these organizations, as the customer, it is imperative that they remain in control of their own strategy with focus on their core competencies. Organizations are getting to be demanding about what service level is guaranteed from the providers (to whom mobile computing solution management is outsourced). This way, organizations do not need to worry that do not have the technical skills to maintain the solution, as long as they know exactly how much this solution—which is backed by a service-level guarantee and has high levels of security—will cost them.

MARKET PERSPECTIVE FOR MOBILE COMPUTING AND M-COMMERCE

With ever-increasing sales of mobile computers, handheld PDAs, and now a new breed of Web appliances and smart telephones, mobile computing has been hailed as a hot new technology that will significantly change the way in which we conduct our work as well as non-work-related activities. Many vertical industries, such as financial services, public safety, healthcare, and utilities have adopted mobile applications since the early 1990s. Even horizontal applications, such as field-service dispatch and Internet e-mail access, have made significant gains recently. Since 1993-1994, early adopters of mobile computing—UPS, Federal Express, Sears, Xerox, IBM, and Merrill Lynch among them—have demonstrated the potential of the technology. Now mainstream businesses are ready to adopt this technology in a serious fashion.

This section provides a perspective on the journey of mobile computing progress. Major players in this journey have been the various committees for Internet governance and technical issues, the Internet service providers, the mobile handset manufacturers, and so on.

According to one subject matter expert (Mallick, 2003), the history of mobile computing shows that solutions in the 1980s and early 1990s concentrated on pure access such as the use of modem racks and *remote access servers* (RASs). Although these solutions were inherently secure, requiring one-to-one connections between the user and the computer, the cost of sufficient telephone lines to meet anticipated peak demand, and the calling costs for users who needed to connect from abroad made the solution suitable only for the most important of tasks. There were bandwidth constraints too. This led to the rise of companies

providing thin client solutions to existing applications.

Meanwhile, the Internet continued to evolve. With the advent of the Internet, solutions began to consider utilizing this network as a means of providing lower-cost access from around the world, utilizing *points of presence* (PoPs) provided by regional Internet service providers to access low-end hosted e-mail solutions for the exchange of information.

Large telecommunications companies and IT vendors such as BT (British Telecom) and IBM then enabled access from these PoPs to their networks, and enabled secure access back into corporates through the use of *virtual private networks* (VPNs). Again, the business models around these solutions were aimed at large organizations with leased line data solutions and depended on high user payments. Complexity was compounded with the advent of *mobile telephony* solutions driven in Europe by GSM. This technology provided low-speed, high-cost means of connecting through to corporate solutions while disconnected from any fixed network, and necessitated support for new mobile phone devices and connectivity between the phone and the laptop.

With the Internet providing a more accessible solution, *security issues* came to the fore. Driven by technologists, the usability of provided solutions tended to be marginal, and companies using the solution still needed to have plenty of bandwidth themselves to be able to support the needs of multiple users coming in through the firewall to access e-mail, calendaring, and scheduling, along with any other functionality they were trying to access.

Today, the market is demanding complete mobile solutions—at the right price, with the right levels of security, and with low management costs to the using company. These solutions must be flexible, so that the company can concentrate on its core competencies, without

needing to employ expensive resources to track, implement, and maintain changes to the underlying transport technologies.

MOBILE COMPUTING CASE STUDIES

Having discussed the overall scenario on mobility and mobile computing, this section provides illustrations from some real-world reported examples (Longbottom, 2003) of how effective mobile solutions can be utilized to competitive advantage. These illustrations come from diverse scenarios such as Engineering Design, Logistics, Insurance, Sales, Field Service, and Home-Working.

Industrial Automotive Sector

The Business Scenario

A European automobile company has followed the trends within the market to the extent where very few components are now manufactured by the company itself. Component manufacture is now outsourced to the lowest cost provider, within the constraints of quality, and this has led to component manufacture being spread across the globe.

The Challenge

The lifecycle of a style of a car is about three to six months, and the company must be able to change the styling rapidly as the market dictates. Changes to the external styling of a car can force changes to underlying components, or to assemblies of these components. The company found that the need to keep returning to the component manufacturers for new designs, and then looking at how these impacted the assemblies, meant that windows of oppor-

tunity in the market were being missed. The company began to lose market share, yet it did not have the option of bringing manufacturing back onsite.

The Opportunity through Mobile Solution

The company decided on the use of a “virtualized solution,” enabling the suppliers to be brought together using collaborative technologies for a Web-based meeting. Through the use of application sharing, component designs can now be compared and checked, ensuring that any inconsistencies can be rapidly dealt with. Also, specialized skills can be brought in to the meeting, which may change the design requirements.

The Benefits

The use of a *virtualized meeting capability* has enabled the required skills and attendees to share views and knowledge without the need or cost of travel, and the design can be finalized within a matter of hours, rather than days. The company involved has reduced the cost of a redesign from many millions of Euros by at least an order of magnitude, and can now cycle designs in days, rather than months, enabling them to hit the market windows more precisely.

Logistics

The Business Scenario

A pharmaceuticals logistics company had a specific, but intractable problem. Deliveries of high-value pharmaceuticals were being made through standard means—a driver turned up in the morning, took a sheaf of papers, loaded up a van, and drove around delivering the loads.

The Challenge

Everything was carried out on paper, which often got lost or was damaged in use. Several of the companies receiving the deliveries soon discovered that it was possible to report a non-delivery to the logistics company, who had little capability to identify whether this was the case or not. On the rare occasions where signed paperwork could be found, the complaining company could just say that they were at fault, and that they had just found the delivery. This fraud reached a level where the logistics company was heading for bankruptcy. A solution was required, and it was required rapidly.

The Opportunity through Mobile Solution

The company decided to move from a paper-based system to a mobile-enabled system. Within this solution, paper dockets were removed, with everything now being carried out electronically. The driver was now in touch with the central office at all times, and thus itineraries could be changed, extra information provided, or issues raised and dealt with at the point of delivery. The solution consists of a set of communication and collaboration components, integrated into the company's own ERP system and to a geo-location system that tracks the route of the vehicles.

The Benefits

The company involved believes that the solution has not only eliminated the fraud involved, but has improved the efficiency of their logistics, due to fewer non-deliveries and a higher degree of issues being dealt with at the point of delivery due to better information availability.

Insurance

The Business Scenario

A major insurance company carries out most of its business through intermediaries. These intermediaries spend a lot of time on the road meeting with prospects, who they are trying to get to sign up for life and health insurance. If a person is interested, they need to fill in a set of forms. The intermediary sends in these forms to the insurance company, who then provides a quote for the prospect.

The Challenge

This takes about three to four days, during which time the prospect has had time to re-think, and closure rates were low.

The Opportunity through Mobile Solution

The company decided to provide each intermediary with a mobile solution, consisting of a laptop with a suite of tools. Now, the intermediary can fill in the form for the prospect on the laptop, and can immediately send this back to the insurance company for an immediate quote. If the company has any issues, they can interact with the intermediary while they are in the prospect's home, ensuring that the issue is dealt with there and then.

The Benefits

With this solution, the company has persuaded intermediaries to concentrate on selling their product, as the closure rate is far higher than the manual systems still used by other companies, and has significantly increased its share of the market.

Sales Opportunity

The Business Scenario

A salesperson is en route to India for a major presentation to one of the company's biggest clients the next day. His departing flight is delayed and he finds himself sitting at the airport for several hours.

The Challenge and the Opportunity through Mobile Solution

During this downtime he inserts his Wi-Fi card and quickly locates the Wi-Fi hotspot in the airport lounge through the use of a suitable mobile client, such as GRIC's Mobile Office client. Once connected to the network, he downloads e-mail and discovers the client's requirements have changed substantially: to win the business he will have to change his offer considerably. He checks the hosted *Sales Force Automation System* (one of a range of personal and group productivity tools available to him wherever he is in the world) to verify some of the client's background information to make sure the new proposal will be appropriate. He accesses the ERP system back at the center via a VPN and personal firewall, and checks the warehouse for suitable parts and to block them from being sold in the meantime; he can request via e-mail that the logistics manager makes contact to show when delivery can be made.

He downloads the latest confidential pricelist from his company intranet. He then organizes a quick Web meeting with the technical department, the product manager, and the prospect to make sure all new information is understood and that the available parts will meet the prospect's requirements. After his Web con-

ference has ended, he learns his plane will be delayed three hours. He sets up shop in a comfortable place and begins drafting a new presentation and proposal. After a few hours he has finished. He orders another drink, reconnects using his Wi-Fi and mobile access client, and sends the new presentation and proposal back to the product manager for input.

The Benefits

A quick call and they agree on terms. He shuts down his connection, boards the plane, and catches a few hours sleep. Once there, he races to his hotel, quickly freshens up, and connects using a local dial-up connection through the same mobile client. He has received an e-mail with final status on availability of parts and delivery time. When arriving at the meeting, the salesperson feels certain to have all final information and can proceed to give a targeted presentation that shows he knows his customer and does not need to walk away with a list of questions to check out "back at headquarters," thus ensuring that the deal is closed with a grateful customer.

Field Service

The Business Scenario

Field engineers generally receive their day's tasks prior to leaving a depot or home through dial-up data synchronization technologies.

The Challenge

Tasks can often not be carried out due to absence of customer, road and traffic problems, or lack of information needed to solve the issue.

Opportunity through Mobile Solution and the Benefits

Through a suitable mobile solution, the engineer can access mapping and traffic information, ensuring that they can actually access a field site. Should they not have the right information at hand, they can access information from a central source, whether this be technical documents or a knowledge base of other engineers' solutions. Should a specific person resource be required, that resource can be brought in through a Web meeting to share written information. Finally, should the engineer finish a job early or not be able to get to any job, the mobile solution can provide further tasks, enabling the company to minimize response times and maintain customer satisfaction.

Home-Working

The Business Scenario

A company has the need for further resources in dealing with customer requests.

The Challenge

There is a shortage of available space in the existing office for new resources to sit. The choices seem to be either to overcrowd the office or to rent further office space.

Opportunity through Mobile Solution and the Benefits

With the correct mobile solution, the correct resource can be found irrespective of geography. By providing access through to communication and collaborative functions, the new resource can work from home while maintaining contact with the rest of the existing team. Access can be provided to existing applications

within the company, while tools such as instant messaging can be utilized to ensure that problems requiring input from other resources can be rapidly addressed. Also, with suitable reporting in place, central office can ensure that the home-worker is using the systems and is meeting targets.

Mobile B2E Outplaces M-Commerce

Mobile commerce has been the driver of much of the early wireless hype. M-commerce has also been the subject of debates and disputes, given that many consider it to be too early to assess and comment on benefits provided by m-commerce. Dholakia, Dholakia, Lehrer, and Kshetri (n.d.) provide some interesting discussion. According to them, mobile phones, mobile Internet access, and mobile commerce are growing much faster than their fixed counterparts. Several characteristics of mobile networks make them more attractive than fixed networks for less developed countries. Comparative analysis reveals the following:

- Global leaders in land-based telecommunications or Internet access are not necessarily the global leaders in mobile connectivity.
- When it comes to the design of m-commerce applications, multiple designs are likely to coexist and compete for an extended period of time.
- Sources and reasons for national leadership in the evolution of mobile commerce applications are likely to be significantly different from the national leadership patterns on the Internet, landline telecommunications, and computers.

A point to be noted is that this is different from most IT fields, where dominant designs have converged rapidly to one or two stan-

dards. When it comes to standards for mobile computing, a peculiar difficulty faced is that of “variety.” Mobile phones alone have a variety of standards. To summarize the standards issue, listed below are the categories of standards for mobile phones:

- First Generation (1G)
- Second Generation (2G)
- Enhanced Second Generation (2.5G)
- Third Generation (3G)
- Fourth Generation (4G)

The mobile computing scenario for m-commerce is also replete with socio-cultural issues. For example, as reported by Dholakia et al. (n.d.), in the Asian countries where mid-range income is becoming the new market target, the consumers from Middle Class strata, would be satisfied with smaller electronic devices and mobile handsets.

Notwithstanding this, proponents of m-commerce, though, have envisioned a world where consumers shopped happily via wireless Web devices including phones and PDAs. Consumers, particularly in the U.S., have been slow to shift buying patterns as expected. The most notable reasons for this seem to be limited screen size and slow performance.

An older study released in mid-2001 by management consulting firm A.T. Kearney quantifies the change in attitudes as the hype has died down. The Industry Standard summarized the study, publishing the following (Industry Standard 2001):

Researchers surveyed more than 1,600 mobile phone users throughout the U.S., Europe and Asia and found that only 12 percent said they intend to engage in m-commerce transactions. That's down from 32 percent just one year ago. And less than 1 percent has actually made any purchases with their phones in the past year.

The place that mobile technologies are gaining the most traction in the enterprise is with business-to-employee (B2E) applications. Companies are finding cost savings and productivity increases by providing their staff with mobile devices and mobile information access. In a recent article highlighting the results of a META Group study titled “Wireless Adoption, Trends, and Issues,” CyberAtlas (2001) notes the following:

As might be expected, organizations with heavy use of pervasive devices by employees are more aggressive in implementing leading-edge wireless/mobile infrastructure components. Several studies have found that the first priority of implementation is for business-to-employee (B2E) applications, because these applications deliver the most immediate productivity return for organizations.

Issues and Challenges

Although mobility presents us with major opportunities, there are and will continue to be business and technical issues. We must deal with these issues, or at least be aware of them, so that we can maximize the effectiveness of the mobile computing solution. The issues for the mobility market are classified into two major parts: *business issues* and *technical issues*. These issues arise partly due to complexity of technologies, combined with the need to deal with multiple access points and concerns around corporate security and the impact of consumers purchasing handheld devices (Godbole, 2003). Higher-speed data telephony solutions, including GPRS, aided by the growth in Wi-Fi hotspot access points, will combine with the explosive growth in devices such as tablet PCs, smart phones, and hybrid devices to create a highly complex, fast-moving market. Here, the key to success will be in knowing how

to create market differentiation through using the available technologies—rather than having the skills in-house on what the technologies are themselves.

Business issues of mobile commerce tend to rise due to a number of reasons. For example, organizations will not find it easy to dictate the kinds of devices their users will have. According to a white paper on the business case for device management of by SyncML (<http://www.syncml.org>), handheld devices used for mobile commerce can pose haunting and daunting issues. These devices also give rise to security concerns. Outside of specific areas where specialized devices from companies are required, device choice is rapidly becoming predominantly a personal lifestyle choice. Business organizations will need to recognize this and frame appropriate security policies that advise on minimum capabilities that a device should have for the company to provide any level of support for it. This allows usage of mobile computing devices with minimal computing features that are called for within organizational context for mobile commerce. As the Web has become ubiquitous, even small companies have increased their global reach. Whereas it was expected that the technology solutions that emerged along with the Web (e.g., video/teleconferencing, instant messaging) would minimize the need for travel, the amount of international travel continues to increase. Although the face-to-face experience does provide value, one person cannot always have all the answer, and it is here where collaborative and communications technologies can provide the missing pieces of information.

Market forces tend to play a part in a sluggish market, where the cost of setting up full facilities for staff can be prohibitive. The use of home-based workers enabled through a set of virtualized functions can provide fast response to market requirements at minimum

costs. Similarly, resources can be optimized through the use of part- or shared-time workers working from home in the same manner. Overall, to enable the uptake and adoption of remote and mobile tools, there is a need for a single, easy-to-use client on the user's device enabling access not only to central networks, but also for responsive, hosted collaborative functionality. Given this scenario in the mobile computing market, within business, profitable market opportunities are created through one of three ways:

1. eliminating organization costs from fixed network solutions (i.e., the conventional wired network-based access),
2. controlling variable costs within an organization, or
3. creating higher margins through the more effective sale of existing or new goods

Of these, the need to control variable costs is the one with the least risk. However, attempting to control costs, without the changing of business processes and business models, could prove dangerous. Truly controlling variable costs needs a new look at how business is conducted, why it is done that way, and how it can be done differently in today's mobile computing paradigm to facilitate conducting business in the future.

With mobility, the issues must be grasped and resolved in the earliest possible stage before it gets too late to control. Mobility of workforce provides opportunities for organizations in the digital economy to create new business processes, re-engineer some of the non-value-adding business processes with the aim to create higher profit margins, and develop new commerce channels within the extended enterprise while controlling implementation costs, as well as costs for support and maintenance. We also have the opportunity to create

a flexible mobile infrastructure, and the foundation to mobile computing operators for gaining market share from their competitors.

The key to using mobility as a differentiator involves more than just looking inside the firewall; it involves extending thinking to company users outside the firewall. By looking at how mobility can influence and facilitate working differently—not just automating what we already do—we can look at how we can maximize the utilization of our expensive human resources. We can ensure that decisions are made on the right information, provided at the right place, at the right time. In this game, the players include suppliers and customers in our *collaborative work paradigm* in an open, yet secure manner. Organizations need to consider utilization of virtual teams consisting of disparate expertise to resolve issues irrespective of location.

Thus, mobility and m-commerce need to leverage this in the market for profitable differentiation. To make the most of mobility, we need to understand the differences between effectiveness and efficiency. Efficiency is just being able to do more with the same resources—even if what we are doing is wrong. Effectiveness is being able to increase the value provided to the company by doing more and more things correctly. Effectiveness is often governed by access to real-time or near-time information between workers, and those with the highest requirement for this access tend to be the most mobile. For example, task-driven groups such as field engineers can plan their routes more effectively when they have access to the latest information. Sales forces can ensure that customer expectations are accurately set through good knowledge of inventories and supplier lead times. Managers can make the correct strategic decisions while on the move. The entire organization can become more responsive, more effective, and therefore more profitable through collective organizational intelli-

gence enabled by the correct utilization of mobility across the whole organization.

As part of this effectiveness, we also need to look at how we use resources for knowledge sharing, utilization, and knowledge management—often we have in-house skills that we only occasionally utilize. If we can provide a flexible collaborative infrastructure enabling resources to exchange views and information irrespective of geography, we can optimize the use of these resources through making their skills more globally available, or outsource certain areas of skill so paying only when these skills are utilized.

When it comes to the implementation of a mobile infrastructure as an opportunity to outsource the solution, it is unlikely that 100% outsourcing of all existing solutions may happen at the same time. Therefore, it is important that any mobile solution provide the tools required for connecting back into organizations' central networks, so that existing applications can be accessed by users through appropriate means.

Consider, for example, secure VPN connection, driven through a non-technical, user-friendly front end. Again, although providing access through to existing applications can provide short-term benefit, one needs also to look at how to be able to evolve the solution through the use of emerging technologies to be more effective in the mobile context. The most important of these emerging technologies will be *Web services*, where applications will be able to make their functional components available to other applications through a series of standardized calls. For outsourced solutions, the emergence of Web services provides a means of enabling information to be interchanged between the managed solution and the existing internal applications. There are many vendors in the market (e.g., Attachmate, IBM) looking at how to take existing "legacy" applications and enable them for Web services.

Although older mobility solutions used *direct-dial solutions* with relatively high levels of security, the very nature of today's mobile needs predicates the use of the public Internet. Therefore, security will be at the top of everyone's concerns. For mobile solutions, the use of VPNs provides a secure means of "tunneling" through the public Internet. However, additional security can be obtained through the use of additional encryption, and external attacks to the client can be prevented through the use of personal firewall technologies. As the solution will also be enabling users to connect back into your own environment, you should consider the possibility of attempts to hijack the connection, and look for mobile solutions that guard against Trojans, Piggy-backs, Worms, and so forth.

Outsourcing is the current vogue. However, many organizations worry about "letting go" of their solution. Often, their concerns are around the inherent dangers with letting an outsourcing company take control of hardware, application, and data. There are several factors involved in this. Firstly, it is likely that any outsourcer will have better physical security than can be provided in-house by most companies. Secondly, the outsourcing company is likely to positively vet its employees—again, an area where few organizations have the capability to do other than cursory checks. Thirdly, the skills that the outsourcer employs will be dedicated skills, ensuring higher availability and better current knowledge of technology than can be afforded within the company itself.

It must also be appreciated that "technology," "data," and "function" should not be utilized interchangeably. Technology is essentially a set of plumbing which can easily be outsourced. Data can reside anywhere and can be accessed by the technology. Function is what the technology and the data support, and is what the company and the end user must have.

Mobility: Opportunities and Issues to Organizations

As per Quocirca reports (2004a, 2004b), the main barriers to entry into m-commerce are *security*, *performance*, and *skills scarcity* (i.e., the need for specific skills to manage a rapidly changing technology environment); and in most mobile computing-related surveys, the main concerns of people are how to open up their own systems to provide this access, without compromising on security, performance, and system availability.

Thus, the inclination seems to be towards use of third-party, managed, hosted solutions.

ALLUDING TO LEGAL CONSEQUENCES OF MOBILITY IN BUSINESS

This turns out to be a serious issue in today's mobility and mobile computing paradigm. Legal aspects of mobile agents, given their increasing deployment in mobile computing solutions for m-commerce, become entities for attention. "Agents" are a special form of (or improvement on) components. They are goal directed and usually self-activating in the sense that they possess their own thread of execution and can initiate actions without intervention from the outside (incoming communication, user interface, etc.). An important distinction between agents and components (on which they are usually based for implementation) is that agents can decide whether to fulfill requests made to them, while components just execute commands.

Mobile agents can travel from one host to another, taking their code, data, and state with them. Because of this mobility, a number of legal problems can arise. Stationary agents, which always remain on the server on which

they were created, are less of an issue: they remain completely under the control of their owner if locally created. There are many agencies to which a finger can be pointed in case of disputes arising out of malfunctioning of these mobile agents, which could result in loss to the customer. Potential entities for this blame are the network provider, programmer of the agent's code, owner of the mobile agent, and owner of the server. Sonntag (n.d.) provides interesting discussion on the topic of usage of agents.

Certain open issues come up:

- What are the legal consequences of mobile agents traveling through different countries? In contrast to ordinary Internet traffic, they are not just transported, but probably executed on servers in other countries. If they possess some information, which is illegal in this country but lawful in the countries of its origin and destination, could the agent be stopped? What actions are required with this information to constitute an offense, or is the mere existence sufficient?
- The technical side is even more challenging. It is often difficult to identify the server, which is responsible for the destruction or a certain change in an agent. Even if extensive logs are used, following them is extremely expensive and difficult, especially if they are located in other countries (often a court order is required for disclosure).
- In some cases electronic signatures and logs would allow automatic arbitration with the evidence collected. However, what should be the result: a fine, an e-mail to the parties? How would it be enforced?
- Similar to users on the Web, the nationality of an agent is easily defined: its owner's nationality. But how can this be detected and/or verified? One possibility is digital

certificates, but the nationality is usually not verified and included in them. This is especially important for blocking or disallowing some actions for agents of a certain nationality, for instance, to enforce taxes or not to sell certain goods to some countries. A solution could be using attribute certificates certifying the nationality, similar to the authorization for signing for legal persons.

CONCLUSION

In essence, m-commerce is the convergence of wireless communication technology and Internet-based e-commerce. It can be expected that m-commerce will be more attractive to customers (or online users). M-commerce has two distinctive advantages over brick-and-mortar business as well as over e-commerce, namely flexibility and ubiquity. Through wired networks, consumers with mobile handsets can conduct business transactions without being fixed at a computer terminal. And as the number of mobile phone users is much more compared to the number of Internet users, businesses can expect to reach a larger market. For the information commerce sector, m-commerce also offers personalization and real-time speed. While the Internet provides a huge amount of information on virtually any topic, mobile devices can disseminate information.

FUTURE DIRECTIONS

With m-commerce and mobile computing applications, users are able to customize and personalize their mobile devices to retrieve information specific to their needs. The speed in which m-commerce transactions can take place is also worth noting. Consumers typically receive

almost instant response for information-related services. Compared to wireless voice or the Internet alone, the wireless Web offers to its users greater convenience. For example, its users do not have to waste their time looking up phone numbers whenever they want to place business orders. Database access for company listings, product availability, and order information can be readily available on the wireless phone. Searching for competitive prices when shopping can be more conveniently done through the Web.

However, m-commerce is still at its infancy. The most popular applications so far have been for entertainment and information-related services, such as the instant messaging service. Although wireless billing is a promising practical application, the security issue must be tackled carefully for this application to be viable. Early implementations of m-commerce have been extensions of e-commerce sites. Technologies such as WAP and Wireless Mark-up Language (WML) enable just that. For businesses, this would translate into better communication with their customers, unlimited by trading hours and geographical distance.

Although m-commerce merely serves as an extension to e-commerce sites, it has been regarded as a value-added service: offering specific information, entertainment, or a transaction over public or private mobile telecommunications networks. According to *Worldwide Market Analysis and Strategic Outlook 2003-2009* (2004), there are certain industry estimates suggesting that between 2002 and 2005, 600 million Internet-enabled mobile phones will be sold. As for revenue, it turns out that the European m-commerce market was \$24 billion by 2003, while end users' spending on mobile e-commerce services were expected to reach more than \$200 billion in 2005. These figures show that m-commerce has many positive value-added services to offer in a huge market. This

indicates that with the high likelihood of m-commerce reaching more than 600 million by the end of 2006, businesses cannot afford not to pay attention to this new medium.

In the future, when m-commerce security is more robust, the use of mobile devices as digital wallets will become more popular. This would enable consumers to shop and pay using mobile devices much like a credit card, as the credit information is stored in those mobile devices. Future and potential applications of m-commerce will involve "information-rich" products and applications with multimedia capability.

A few more concluding remarks need to be in place to end this chapter. Mobility is on the rise—it is important to see it as an opportunity, not a threat. Many organizations may look at mobility as purely an opportunity to extend their existing systems—those who take the opportunity to change business processes and look to mobility as a means to an end will be the winners. Also mobility is not just about connectivity—it is about function. Further, it is important to understand that MOBILITY can be highly technical, so organizations must concentrate on making it easy for the user. If the solution is seen as being of little complexity to the user and is couched in terms that make business sense to them, they will use it. There are many issues and challenges: mobility and mobile computing is replete with its own business and technical challenges.

Today belongs to "business-driven IT." That means it is important to keep in mind that IT is not a core competency. Thus, organizations are looking for outsourced solutions to provide a more flexible system that should be more responsive, available, and manageable than an in-house solution. Also, in-house is constraining. Mobile technologies are changing—and putting in place an in-house solution will require continuous change in the coming years. This will adversely impact organizations' companies

capabilities to compete and will incur heavy costs. This will further enhance the trend towards outsourced solution because that can provide a more flexible environment.

REFERENCES

- CapGemini. (2001). *Internet mobile for business: So far from expectations*. Retrieved April 18, 2005, from <http://www.capgemini.com/news/2001/0618mobile.shtml>
- Cellular. (2005). 'Third generation' mobile technology 3-G key and WAP glossary. Retrieved May 25, 2005, from <http://www.cellular.co.za/technologies/3g/3g.htm> and October 7, 2005, from http://www.refreq.com/WAPTech/wap_glossary.htm
- CyberAtlas. (2001, September). Business-critical applications driving wireless initiatives. *CyberAtlas*.
- Dholakia, N., Dholakia, R. R., Lehrer, M., & Kshetri, N. (n.d.). *Patterns, opportunities, and challenges in the emerging global m-commerce landscape*. College of Business Administration, University of Rhode Island, Kingston, USA.
- Godbole, N. (2003). Mobile computing: Security issues in hand-held devices. *Proceedings of the NASONES 2003 National Seminar on Networking and E-Security*.
- Godbole, N., & Unhelkar, B. (2003, December). Enhancing quality of mobile applications through modeling. *Proceedings of the Computer Society of India Convention, IIT-Delhi, India*.
- Hansmann, U., Merk, L., Nicklous, M., & Stober, T. (2003). *Principles of mobile computing*. New Delhi, India: Springer.
- Industry Standard. (2001, May). Analysts change their tune on m-commerce. *Industry Standard*.
- Intel. (n.d.). *Intel Technology Journal, archival of past journals*. Retrieved from http://intel.com/technology/itj/q22000/articles/art_6.htm
- Jgohil.Typepad. (2005, January 18). *A case on NTT DoCoMo*. Retrieved from <http://jgohil.typepad.com/minternet360/imode/>
- Kerner, S. M. (2005). *PDA market up or down?* Retrieved May 23, 2005, from <http://www.internetnews.com/stats/article>
- Kerner, S. M. (2005, July 22). *Cell phones rising*. Retrieved May 19, 2005, from <http://www.internetnews.com/stats/article.php/3522076>
- Kolari, J., Laakko, T., Hitunen, T., Ikonen, V., Kulju, M., Suihkonen, R., Hu, S., Virtaneen, T., & Tytti. (2005). *Context aware services for mobile users: Technology and user experiences*. Retrieved May 20, 2005, from <http://www.vitt.fi/inf/pdf>
- Longbottom, C. (2003). *Mobility as an opportunity*. Report commissioned by GRIC.
- Ly, F., & Sugianto, S. S. (n.d.). Business models in the digital economy. Retrieved April 15, 2005, from http://aisel.isworld.org/subject_by_publication.asp?Subject_ID=17
- Mallick, M. (2003). *Mobile and wireless design essentials*. New Delhi, India: Wiley-Dreamtech India.
- Mobile Applications Initiative. (n.d.). Retrieved from <http://www.mobileapplicationsinitiative.com/>
- MobileInfo. (2005). *One stop Web site for mobile computing and wireless information*

to access hot topics in the mobile computing industry. Retrieved May 25, 2005, from http://www.mobileinfo.com/3G/3G_Wireless.htm

Quocirca. (2004a, Summer). Enterprise wireless update. *Quocirca Insight Report*.

Quocirca. (2004b, November). Optimizing the mobile workforce—A business value perspective. *Quocirca Insight Report*, 6.

Quocirca. (2005). *Mobile devices and users*. Retrieved April 18, 2005, from www.quocirca.com

Reiter's. (2005, June 6). *Reiter's camera phone report*. Retrieved April 18, 2005, from http://www.wirelessmoment.com/statistics_camera_phones/

Sonntag, M. (n.d.). *Legal aspects of mobile agents with special consideration of the*

proposed Austrian e-commerce law (pp. 2, 3). Institute for Information Processing and Microprocessor Technology (FIM).

SyncML Initiative. (n.d.). *Business case for device management*. White Paper. Retrieved May 25 and October 7, 2005, from <http://www.syncml.org>

Telecommunications Service Inquiry. (2003). *Data and mobile services. 2000—Connecting Australia*. Commonwealth Department of Communications, Information Technology and the Arts. Retrieved May 20, 2004, from <http://www.teleinquiry.gov.au>

Worldwide Market Analysis and Strategic Outlook 2003-2009. (2004, August). Future mobile computing: Device trends and wireless solutions. *Worldwide Market Analysis and Strategic Outlook 2003-2009*.

Chapter XXXIV

The Future of Mobile Technologies and Applications in China

Xiao Chen

Nanjing University of Chinese Medicine, China

Wei Liu

Nanjing University of Chinese Medicine, China

ABSTRACT

This chapter deals with the future of mobile technologies and applications in China. The effect of emerging technologies, especially mobile technologies, on the massive market of China cannot be ignored in the global context. This chapter gives the reader an insight into China's mobile telecommunication industry today. The authors firstly relate statistics about China's mobile business market including user and device analysis that helps in providing an understanding of mobile business in China. This analysis is followed by a description of the major mobile technologies employed in China and a brief view of the Chinese market's status, followed by an insight into some newly rising industries which are potentially successful mobile sectors in China. Finally, a real life example is examined—that of M-Government Project in Gunagzhou, capital city of Guangdong Province.

INTRODUCTION

After the arrival of the Internet in China in 1987, there were huge changes to the way the Internet developed further in China. Earlier, without the popularity of the Internet, both people and businesses used dial-up modems to log onto the Internet. Later, Digital Subscriber Line (DSL) technology came into common

people's lives. A few years later, it was the application of broadband, which made the connectivity to the Internet very popular. Finally, in 2004, wireless technology further extended the application of the Internet in people's lives as well as in businesses. These various methods cover almost all markets of Internet access; they are both beneficial for Internet service providers (ISP) and Internet operators. They

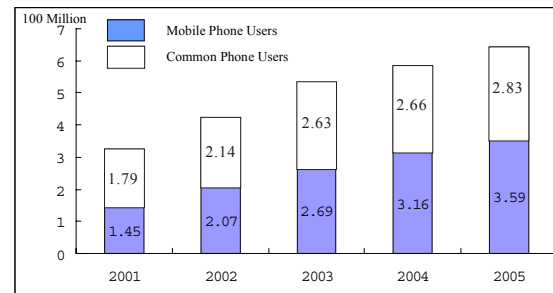
bring in plenty of profits for the two major Internet operators—China Telecom, and China Netcom—and the two make the Internet market more ubiquitous. However, traditional access methods like dial-up and DSL technology need stable network facilities for the support, therefore it will not be convenient for the users to use the Internet without distance limit, and it might be a limit for business use and ensure the business processing in anywhere customers needs. For the high demand of moveable and convenient business, a new-generation connecting method called “Mobile Internet” (Wireless Internet) has been developed; its the main technologies are wireless application protocol (WAP), General Packet Radio Service (GPRS), Code Division Multiple Access (CDMA), Centrino, and Chinese Wireless Authentication and Privacy Infrastructure (WAPI). WAP was the first trial of Mobile Internet in China, but it seemed to be a failure; Centrino and WAPI are the two latest trends of Mobile Internet in China; both have advantages and disadvantages. Which will be chosen by the future, and which will be dominant in China’s Mobile Internet industry? Let us take a look. In this chapter, we discuss the importance of mobile technologies and applications in this modern era with relevance to mobility in China.

MOBILE STATISTICS IN CHINA

Business Statistics

The information industry has been the one of the major drivers to impact economic development in China. And within the information industry, the communication and network markets have continued to maintain high-speed progress. Up until the end of 2003, the number of telecommunication users in China exceeded 400 million, and the number of Internet users

Figure 1. Trend of increment of China’s common phone users and mobile phone users from 2001 to 2004, and forecast for the number in 2005 (CLI4 Report, 2003)



reached 79.5 million. These figures establish the foundation of the development of mobile technologies and applications in China (OECD Report, 2004).

In China’s domestic telecommunications market, mobile communication equipment and services are widely accepted and adopted not only by individuals but also organizations. At present, mainstream mobile communication equipment includes mobile phones and PHSs (Personal Handphone Systems). As the largest mobile phone market, China had 269 million users in 2003, which increased 30% from 2002. Figure 1 shows the trend of increment of China’s common phone users and mobile phone users from 2001 to 2004, and it makes forecast for the number in 2005 (CLI4 Report, 2003).

Along with the upgrade of technologies, the number of PHS users also goes up dramatically from 2002. In September 2002, the number of PHS users exceeded 10 million, and at the end of 2002, the number increased to 11.13 million. At present, this number continues to increase at high speed (CLI4 Report, 2003). From Table 1, we find the number of PHS users in 2002 was over 11 million, and in 2003 the number boosted to 24 million, double that of 2002. Also we find the numbers of mobile phone users and other

The Future of Mobile Technologies and Applications in China

Table 1. Development trend of different kinds of China telecommunication users (CLI4 Report, 2003)

USER	UNIT	YEAR 2002	YEAR 2003
Common Phone Users	100 million	2.14	2.63
PHS Users	10,000	1,113	2,400
Message Users	10,000	14	45
Broadband Network Users	10,000	357	580
Defense Data Service Users	10,000	46	45
FR and ATM Users	1	9.9	12.6
Internet Users	10,000	4,970	9,200
Mobile Phone Users	100 million	2.07	2.69

Table 2. Capability of China Mobile and China Unicom from 2003 to 2005 (Broadcom Consulting, 2004)

	YEAR 2003	YEAR 2004	YEAR 2005
China Mobile GSM (10,000)	21100.0	23600.0	24000.0
China Unicom GSM (10,000)	8000.0	9000.0	9000.0
China Unicom CDMA (10,000)	3900.0	5000.0	6500.0

kinds of communication users increased by geometric series. Typically, we can see that Internet users in China increased so much that the number reached 92 million in 2003.

After the split and re-unification of the department of China's information industry at the beginning of 2002, two organizations were formed to compete for the mobile communication operation market. These are China Mobile and China Unicom. China Mobile owns the largest Global System for Mobile communication (GSM) and General Packet Radio Service network in China, which covers all the provinces and cities in China. Nevertheless, China Unicom established more than 300 subsidiary companies in 31 provinces in China, and its GSM network also spans most of China, excepting some parts of Tibet and certain villages. Meanwhile, China Unicom also finished its CDMA network, thus China Unicom owns two different networks which both cover the whole China. After the first phase of the project, the capability of the new operation network

reached 15.81 million people, and there are 188 large switches, 9,000 direct stations, and 14,458 base stations. Nowadays, CDMA covers 92% of all of China's 332 cities and is affiliated with more than 1,800 villages and towns.

Similarly, China Mobile invested RMB 300 million for the 3G communication network GPRS in 16 China provinces whose capability was 414,000 users in 2001. Up until the end of 2002, the total capability of GPRS was three million, covering 241 major cities in China. The income of GPRS users comes to more than RMB 80 million.

In the network market, according to relevant data statistic, the global sale income of Wireless Local Area Network (WLAN) was \$7 billion in 2003, and it will increase to \$44 billion in 2008, which means a 44% hybrid increment ratio per year. Although the start-up of China's market is a little late, the average increment ratio is up to 51% per year. In 2003, the total sale income of WLAN in China was \$58.06 million. Furthermore, as a result of

Table 3. Advertisement investment for WLAN in China from 2001 to 2003 (CNIIC, 2003)

YEAR	2001	2002	2003
Cost (RMB 10,000)	283.8	468.2	1547.8
Increment Ratio	---	64.98%	230.59%

coming to an agreement about a WAPI standard with the U.S., the scale of China's wireless broadband network market is two times as large as it was at the end of 2003.

The WLAN operation in China also drives the market of networking access equipment sale. Thus, because of earlier participation, foreign providers could provide a wireless access point, wireless router, wireless bridge, and other system equipments. For instance, Airconnect of 3Com, Aironet series of Cisco, Intel PRO/Wireless 2011LAN, and Intel PRO/Wireless 5000LAN could meet the needs of individuals and different sizes of organizations. Also, domestic manufacturers such as Huawei, ZTE, and DTT start to focus on the WLAN market. They finish their own products in succession to supply to customers; these products include an inner wireless network interface card (WNIC), a wireless client adapter, a solution scheme, and so on.

Also, manufacturers increase their advertisement investment in order to build up visualization and occupy the WLAN equipment market. The

advertisement costs in China from 2001 to 2003 for WLAN products are shown in Table 3.

In addition, the new technology standard also becomes the force to make the WLAN market develop fast in China. Since IEEE802.11 is executed in due form, WLAN technologies grow fast and become mature gradually. Figure 2 lists several standards that are accepted and used in the WLAN market (TechTarget, 2005).

At present, these standard and technologies are widely used in all manufacturers' products. These factors improve the development of China's WLAN market at a high speed.

MOBILE USER ANALYSIS IN CHINA

Generally, the mobile user in China could simply be divided into two parts. One part is the mobile communication equipment user, which means mobile phone and PHS consumers. There is no doubt that the users of mobile phones and PHSs will be continuously increasing. The call service will keep developing quickly. However, we want to introduce and analyze the mobile users who manage data transmission and Internet surfing.

High Development Speed of Data Transmission Operations

In China's mobile communication market, the data transmission operations contain Short Message Service (SMS) based on GSM and Multimedia Message Service (MMS) based on GPRS. Table 4 shows the income of China's telecommunication industry from 2000 to 2002.

Figure 2. IEEE wireless speed model (CNNIC, 2003)

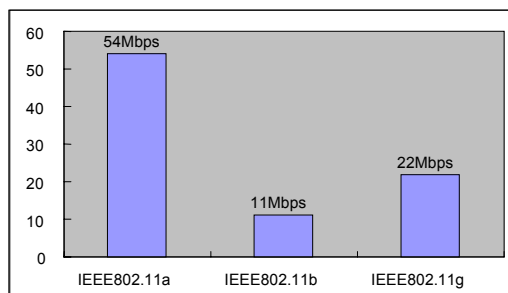
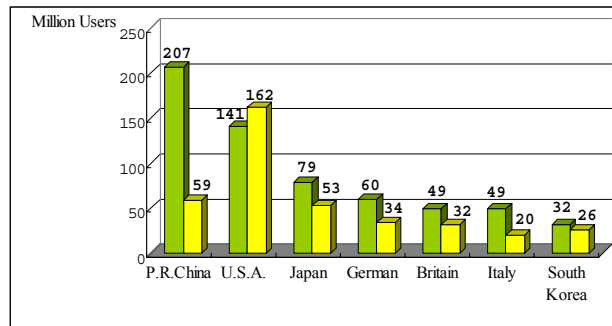


Table 4. Major telecommunication operations income from 2000 to 2002 (CNIIC, 2003)

OPERATION TERM (100 MILLION)	YEAR 2000	YEAR 2001	YEAR 2002
Telecommunication Total Incomes	3006.5	3535.2	4115.82
Long-Distance Communication Income	618.3	528.7	563.8
Ratio	20.56%	14.96%	13.7%
Data Communication Income	72	131	181.1
Ratio	2.39%	3.71%	4.4%
Local Communication Income	928	1027.7	1212.3
Ratio	30.87%	29.15%	29.5%
Mobile Communication Income	1253.4	1617	1954.4
Ratio	41.69%	45.74%	47.5%
Other Operation Income	70.8	181.1	163.02
Ratio	2.35%	5.12%	3.9%

Figure 3. Rate of Internet users and mobile phone users in several countries (CNIIC, 2003)



In 2002, the income of data communication was RMB 18.11 billion, an increase of 38.1% from 2001. Moreover, the percentage of data communication in the total income of China's telecommunication industry rises quickly. Though the proportion is still not very large, the potential increment and development is imponderable.

Improvement of Combination with Internet and Mobile Phone

In 2003, the rate of mobile phone users and Internet users was 3.5:1 in China; no other market achieves this rate (see Figure 3).

Only according to quantity, interest, and development trend is it possible for China to achieve advantage in the field of the combination of mobile phone and Internet.

The manner than young people use China's Internet may indicate the trend of manner for all people. In China, 70% of Internet users are under 30 years old.

The combination of entertainment value, fashionableness, and low price of the mobile gadgets provides an excellent opportunity to cater to the younger demographics of China. Web sites usually provide mobile phone users with the services listed in Table 5.

However, to these young people, the price of the combination of mobile phone and Internet

Figure 4. Age analysis of China's Internet users (CLI4, 2003)

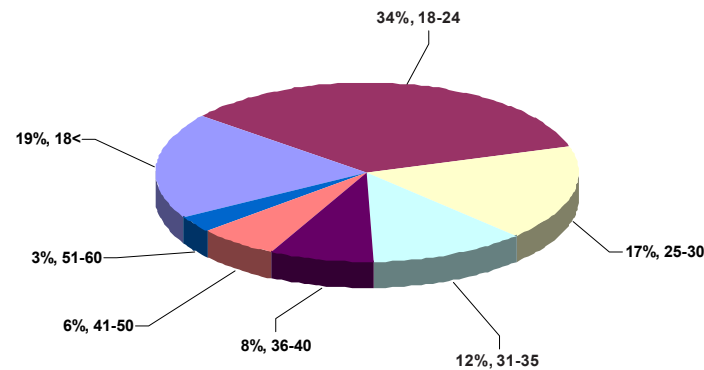


Table 5. Major operation terms that Web sites provide (CNNIC, 2004)

PRODUCT	DESCRIPTION
SMS	Elementary words short message
MMS	Photos/graphic short message
WAP	Online scanning (include searching)
IVR	Obtain voice message and entertainment

is acceptable nowadays. For example, the phone ring and game download, appointment, and short message meet the needs of young people and have large data communication traffic, but the price is not so high. For the major Web sites in China, the income flow related to short message and game starts to increase (CNNIC, 2004).

Also important is wireless network users, including wireless networking access equipment users and WLAN users.

Enterprise Users Increase Sharply and Individual Users Keep Steady Increment

The mobility of WLAN equipment could bring users convenience and higher access speed. Along with the price decline of WLAN equipment such as Access Point and NIC, the number of China's WLAN users will mushroom.

Moreover, as the terminal unit of WLAN equipment, most notebook computers begin to install WLAN NIC. From Figure 6 we can clearly see sales information of notebook computers during the first three quarters in 2003 and 2004.

It is not difficult for us to find that enterprise users provide the major purchasing power, which is about two-thirds of total sales num-

Figure 5. Internet major Web site income category, fourth quarter 2003 (CNNIC, 2004)

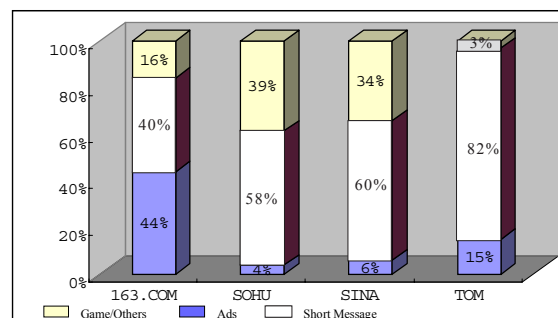


Figure 6. Notebook sale amount of first three quarters in 2003 and 2004 in China (CNNIC, 2004)

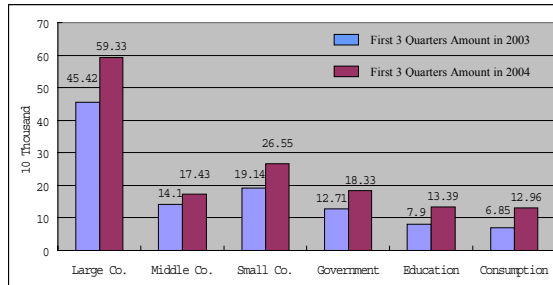
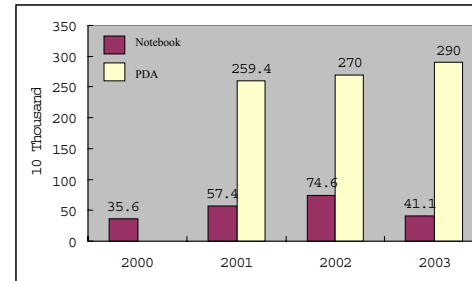


Figure 7. PDA and notebook sales from 2000 to 2003 (Blogchina, 2004)



bers. As more and more districts are covered by the WLAN network, there is no doubt that in the increment of WLAN users, enterprise users will occupy a sizeable percentage. At the end of 2002, there were 25,000 WLAN users: 15,000 enterprise users and 10,000 individual users. As predicted, the scale for enterprise application will exceed the application of the public operation market in 2005.

Industry Users Are the Power for WLAN Market Development

At the present time, WLAN application focuses on public service, campus network, and family and individual fields, but as application use broadens, the focus of the market will gradually transfer to industry application. According to the experience of other countries, the aim for WLAN industry application is high-level business managers. However, the final goal of the WLAN market is to make it popular;

therefore, industry users are the largest group of potential WLAN users.

Potential User Market Is Quite Huge

As the terminal units of WLAN equipment progress, the sale of PDAs and notebook computers is increasing constantly. As one professional comments, the notebook computer is the most suitable terminal unit for the WLAN application. Meanwhile, sales numbers of notebook computers indicate that the notebook is one of the products whose sales are increasing at high speed (see Figure 7).

Up to the second quarter of 2003, the sales number of PDAs was 2.9 million; thus, the number of users who use PDAs to access the Internet is only 8% of all Internet users. Also, in the same timeframe, about 17.6% of notebook users become WLAN users, while among PDA users who use PDAs to connect to the Internet, there are 30% WLAN users.

Table 6. Movement of different WLAN application ratio markets in China from 2002 to 2005

	Public Service Market	Industry Application Market	Family Application Market
2002	42.2%	24.6%	7.6%
2005	12.4%	39.2%	18.9%

USAGE OF MOBILE TECHNOLOGIES IN CHINA

WAP (Wireless Application Protocol), Maybe Too Good for China

Dating back to 1999, the first sales of wireless data services were provided by former China Tele (China Tele and China Mobile Tele today), who shared connected mobile network and paging with Chinanet, the largest Internet infrastructure for business use in China. This service, however, was not balance-developed and could not be served in some provinces in far-away China. Therefore, the customers of this service were a minority, and few people knew about it. With the high-speed growing economy, Internet technology, and mobile telephony in China, a new application was introduced to mainland China—wireless application protocol (WAP). WAP made more and more people accustomed to the new concept of accessing the Internet wirelessly via mobile phones. In 1999, the hotspot for the Internet was WAP. Both operators and equipment suppliers such as Ericsson, Nokia, and Motorola strived to create and provide the trial service of WAP. In March 2000, China Mobile launched its trial service of WAP in six major cities in China, including Beijing, Shanghai, Shenzhen, Tianjin, Guangzhou, and Hang Zhou; meanwhile, China Unicom launched the same service in Guangdong and Shanghai, the two fastest developing provinces in China (Xiangdong, 2000). WAP was finally abandoned by the market due to low speed of transmission, lack of sufficient content and fewer competitors, and high cover charges. The two major operators quickly found that only new, effective technology could satisfy the needs of consumers.

GPRS and CDMA: The Right Choice and Rising Star in China

Eventually, China Mobile and Unicom both found that WAP was not their final product and the correct choice for the m-Internet market. They knew that WAP was the “DOS” of the wireless technology world, and the “Windows” operating platform had not yet been established. GPRS technology is one of two developments chosen by China Mobile after the unsatisfied WAP services in China. The main benefits of GPRS are that it reserves radio resources only when there is data to send, and it reduces reliance on traditional circuit-switched network elements. The increased functionality of GPRS will decrease the incremental cost to provide data services, an occurrence that will, in turn, increase the penetration of data services among consumer and business users. In addition, GPRS will allow improved quality of data services as measured in terms of reliability, response time, and features supported. The only way to succeed in m-Internet is to find the right model for it. Both of these attempt some development.

The other mobile operator, Unicom, chose a totally different service called CDMA, which is a large, successful business in South Korea. With the slogan of “Green Telephone, Less Radiation,” and with the marketing strategy of “less pay, more calls,” in the years that followed—especially in the year 2001—millions of mobile phone users were attached to Unicom as the newest innovation of the m-Internet industry. Another reason Unicom chose CDMA was that Unicom was the sore licensed operator who could provide CDMA service in mainland China. In its long-term blueprint, Unicom deployed its CDMA network in more than 200 Chinese cities, with network capacity of up to

about 15 million users within one year. In 2001, China Unicom updated its IS-95 CDMA to CDMA2001x, s CDMA2001x services, which was the latest CDMA technology worldwide at that time (Xiangdong, 2000).

To this point, we have discussed Mobile Internet service via mobile phone. However, many people like to surf the Internet via computers, especially laptop computers; yet, they prefer to get rid of the limit of wired connection and become truly free. The following technology will focus on the wireless technologies for computer use.

The Newcomer from the U.S.—Intel Centrino Technology: Great Waver in the Mobile Computer Market

In 2002, Intel, the largest microprocessor manufacturer, introduced a new wireless technology called Centrino, which integrates the wireless connection function into the Pentium CPU. This technology is based on IEEE standard 802.11b, and it is proven that Centrino can make the real wireless Internet available without help from a CID card. Also, general performance of PCs seemed to be upgraded, therefore customers gained a more comfortable and multifunctional laptop computer. When first introduced into the China market, Centrino was quickly adopted by many major laptop computer manufacturers like IBM, HP, TOSHIBA, and some other domestic computer manufacturers, like Tsinghua, Founder, and Lenovo. The promotion stage was also done very well, so that Centrino is the representation of wireless technology, ensuring the availability of real wireless technology. Three years into the use of Centrino, users have calmed down and are ready to verify whether Centrino is really “precious” or “rubbish” for the wireless world. Some people say Centrino still needs much support, like a connection base, access portal,

and so on. Furthermore, because the infrastructure construction in China is ongoing, users cannot access the Internet at any place they want, so it cannot be said that the use of Centrino can make users free of distance. In this case, the road ahead for Centrino is still so long that timeframes cannot be determined.

Security Issue: China’s Own WAPI

WAPI is aimed at device manufacturers around the world hoping to produce mobile-ready goods for sale in China, and is a competitor to the Wi-Fi security standard. According to a notice issued by the Standardization Administration of China, which manages standards in various industries, as of December 1, 2001, Chinese government agencies prohibited the import, manufacture, and sale of Wi-Fi gear that did not use China’s new security specification (Nan, 2001).

Presently, all Wi-Fi gear makers are to start using the Wired Authentication and Privacy Infrastructure specification, support for which is not included in current or upcoming security specifications, such as Wi-Fi Protected Access or 802.11i, developed and enforced by the Institute of Electrical and Electronics Engineers and the Wi-Fi Alliance, effectively making it incompatible with the standards developed by technology industry groups. WAPI is to be used with Wi-Fi standards in the 2.4GHz radio band, according to the standards, adding confusion to the market by presenting another obstacle for manufacturers looking to sell products in China, as WAPI adds another security specification that needs to be considered as companies start to install Wi-Fi networks.

Security experts say that, by prohibiting gear that does not use WAPI, the Chinese government is throwing an obstacle in the way of manufacturers looking to enter the Chinese market. “It would be unfortunate if we are not

able to resolve this so that the China standard and others in the world can't coexist," said Dennis Eaton, chairman of the Wi-Fi Alliance. "Wi-Fi vendors may have to use special requirements for products that sell into China." According to Eaton, it is clear that WAPI does not use Advanced Encryption Standard (AES), a key encryption component for wireless networks.

RIISING BUSINESS IN CHINA: EDUCATION AND MOBILE

Together with the high-speed development of the SMS market in China, a new concept of mobile education (m-education)—which combines today's hi-tech communication technologies and new patterns of education—has come out in this continuous changing time. Here we are going to analyze this interesting topic in the range of market possibility analysis, fundamental frame of the m-education, and common applications of the m-education in China. First, let us look at the market analysis.

Market Analysis for M-Education in China

As discussed in pervious sections, the volume of transactions processed in the terms of SMS was huge in the year 2003 in China; it is a definite to say that people in China grow accustomed to using SMS as one of their common communication channels to gather useful information they need. However, if we go through various types of SMS business in China, we may find that the majority of those business transactions were entertainment-based and living-information businesses, like entertainment news, weather forecasts, and so on. There are limited services that could provide people with useful information about education and learn-

ing. In this way, there must be a blank for SPs (service providers) to explore this market (Xiangliang, 2002).

Today in China, two main objectives are to improve education quality and continue to expand the volume of education. How high-tech can help China to achieve those two goals is the hot topic in China today. After analyzing four fundamentals of education—teacher, student, teaching content, and teaching media and facilities—people realize that the thing high-tech can do is change teaching media that Chinese people have used for thousands of years—that of "Chalk and Blackboard." In China, correspondence courses have proved the possibility of remote education, and upcoming broadcasting and Internet education is becoming more and more popular in China. In this case, it seems that education in China is becoming more versatile, and education in China is getting beyond the limits of time and geographic location. However, within the competitive environment, the need for lifetime education is also increasing in China: people who are eager to study but have no fixed time need a more moveable platform to gain knowledge and education information. Furthermore, for instance, students who are on vacation, but who need to keep up with school, also need an effective platform to get such information. In the meantime, with a decrease in costs, a high popularization of mobile devices (e.g., cell phones) among students in China has occurred. Furthermore, the capability of networking and computing technologies improved so much that the bandwidth for communication and data processing is enough for a large number of people to query for various information about education in various formats, like SMS and MMS. In such a background, with high integration of mobile technologies and education, m-education in China seems to be blooming; the market will be very interesting and profitable for people to explore.

Basic Formats of Mobile Education in China

There are many different formats of m-education in China; some are categorized by content, some by technology used in mobile learning, and others by objects. Here we focus on technology-based mobile education formats (Huang, 2002).

SMS-Based Mobile Education

In this model, SMS—this well-developed and widely used format—could not only enable the data connection (encoded characters) among users, but also the data communication between user and Internet server. By using this feature, people could enable communication via wireless networks and the Internet; therefore, the interaction between students and the school will be activated as follows:

- a. Officers (administrative side) could send teaching notes and announcements to teachers.
- b. Teachers could send teaching and study notes and announcements to students.
- c. Teachers or officers could read questions which are sent by their students or clients.
- d. Teachers or officers could answer those questions which are sent by their students or clients.
- e. Both teachers/officers and students could query for their own purposes to gather particular information.
- f. Students could use SMS to check marks and credit points they have achieved.

In a word, via SMS, the information flow among students, teachers, and officers will be direct and unblocked, and people who use SMS could achieve the objective of “study any-

where, and at any time.” However, we still need to solve some problems. At first, the server for the connection needs to be enlarged. Then, workable and effective SMS-oriented software should be developed. After that, a customized interface should also be taken into account.

Internet-Access-Based Mobile Education

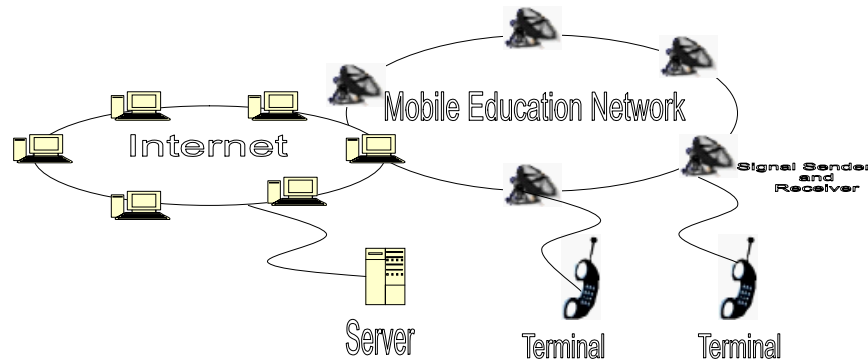
In this model, people could use their multimedia mobile devices to gather educational information from the Internet. Despite the SMS-based model, users could acquire multimedia information by using Web explorers like those installed on their devices (e.g., Internet Explorer and Netscape). With technologies such as WAP, GPRS, CDMA, and even 3G, the speed for wireless connection is become faster and faster, and either the capacity, or capability, or the file format will become larger, more advanced, and compatible. As with SMS-based mobile education, the issues of high-level service fees and software needed to support the service seem to be a long way off for the explorer.

Local Server-Based Mobile

In this model, a mobile workstation is used as its server to provide information sharing among different terminals by using IR or Bluetooth technology. In this case, with a limited number of mobile devices available with IR or Bluetooth features, a WLAN with a limited range could be set up so the data flow among those terminals could be OK, and data and information sharing will be available (Guangzuo, 2004).

Without doubt, the future of mobile learning will flourish; the market of mobile learning will be a big cake waiting for someone who is hungry.

Figure 8. Main fundamental frame of m-education (based on Guangzuo, 2004)



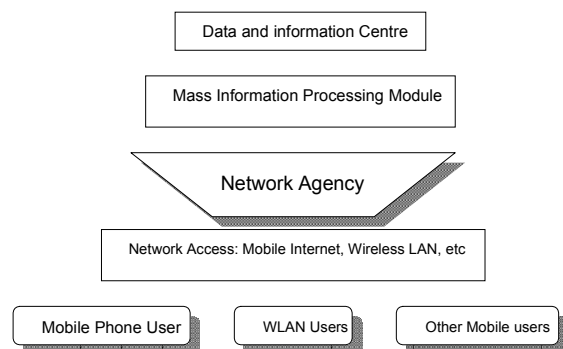
Cases for Mobile Education in China: Peking University's Mobile Virtual Campus (MVC) Project

As the leading university in China, Peking University takes an early step to implement the latest wireless technology as an operating wireless system on their campus; it is called "Mobile Virtual Campus Project." This project was based on the mobile telecommunication and computing technology, and it adopted a distributed infrastructure. As Figure 9 shows, the MVC model consists of five components: Data & Information Center, Mass Information Pro-

cessing Module, General Agency, Networking System, and End Users (Guangzuo, 2002). The Data & Information Center is the main storage device (logical and physical) of MVC, and it is linked to the university's main servers, adding in some open module especially for mobile usability. The Mass Information Processing Module processes all the services required. General Agency is a layer that contains several networking and data communication protocols that enable the users to use MVC. And the users—WLAN users, and PDA and mobile phone users—could use different Network Access Methods to visit and get useful information from MVC. We can conduct MVC mainly based on a top-down model—information was sent from the top, and the end users received the information from their agencies.

Peking University launched this project by implementing 250 wireless routers which covered main buildings like students dorms. Students and teachers could access the intranet and Internet by using their accounts, so that mobile education could become possible. Users such as students could access to their own sites to submit assignments and other academic documents, and teachers could use wireless networking to access the Web to view students'

Figure 9. The MVC infrastructure model (Guangzuo, 2002)



work, and view some teaching materials. From the statistics provided by the Peking University Students Union, there are more than 5,000 MVC users using the mobile networking system, and they have been satisfied by such an effective wireless-based system.

Moreover, the SMS-based platform for mobile phone and PDA communication—called Large-scale Distributed Parallel Mobile (LDPM) platform—is still under construction. It is based on the Mass Information Processing Module, and is especially for the implementation of mobile telecommunication. This system supports various kinds of software developing language like C++ and JAVA, so it is possible to develop various numbers of specific mobile systems on it. Several project research teams have finished the development of SMS-based mobile education systems, including Mobile Question-Answering System, Mobile Office Automation System, Mobile BBS System, Mobile News System, and Mobile Teaching Material Administrative System (Guangzuo, 2002). By the year 2002, the Mobile Office Automation System and Mobile Teaching Material Administrative System were, in their BETA-released version, being used at Peking University. Although some of their functions need to be improved, it is still believed that in the near future, teachers and administrators could edit short messages to send information to their target, and students could use their own mobile phones to get information from the MVC. At such time that MVC is well implemented, both teachers and students at Peking University will enjoy their digital life better.

GOVERNMENT SECTOR AND MOBILE

Whenever talking about electronic commerce or electronic business in China, many people

are very interested in government, and they treat government as an independent but great market which has a lot of potential demands on the information service. In recent years, a new conception called *mobile government* came into being. As we know, the Chinese government has come a long way to achieve an electronic-technology-based working style, and officers have grown accustomed to monitoring and operating their businesses through the computer and the Internet. However, because of the immovability of LANs, which have been deployed in almost every office nationwide, officers and governors are not able to get wanted information if they left their office and without prepared notes. This has caused much inconvenience for the officers and has lowered work efficiency. Those drawbacks make the appearance of mobile government possible.

How can mobile government benefit users? Take a look at a successful case. Guangzhou, capital city of the Guangdong province in China, has launched an information project mainly concerned with how to build a mobile-technology-based platform to run services for the citizens. If the project were successful, people in Guangzhou could use their own mobile terminals to get services from the local government. China Mobile (Guangzhou Branch), leading and largest mobile telecommunication operator in China, was assigned to develop the project. After analyzing the real case, they changed four high-rated government services into a mobile model (Bing, 2004).

1. **Fanyu Project:** Fanyu government used SMS to automatically announce applicants who applied for the permits whether they passed the evaluation and audit process. This service shortened the waiting time for each individual applicant.
2. **Taxation Payment:** In order to get taxation records, a taxpayer could edit SMS

and send it to National Revenue. The system automatically feeds back the enquiry content and provides all related information.

3. **Passport and Visa Enquiry Service:** China Mobile and the Guangzhou Police Department opened service for people to enquire about their passports and Visa records; also, people could apply for a passport by sending SMS to the Immigrant and Emigrant Department of the Guangzhou Police Department.
4. **Custom Declaration Service:** After installed specialized facilities on every ship, custom officers could get geographic, audit, and video data, like location about every ship. This service can also process the custom declaration service by sending particular forms to the server. This service may avoid smuggle affairs and facilitate the processes of Customs.

As we can draw from the above, mobile government could:

- Facilitate processing speed and save a lot of time.
- Save government operating costs.
- Improve citizens' satisfaction.
- Produce a better feeling about local government.
- Improve the decision-making correction level for the government officer.

Besides two newly mobile business models described above, there are still many blank areas awaiting exploration, including mobile business in the health sector, catering sector, and so on. It is easy to foresee that a few years in the future, mobile business could infiltrate into every industry sector in China.

CONCLUSION AND FUTURE DIRECTION

We plan to extend this literature survey and experiment with how China can benefit from use and application of mobile technologies. As we can see from this chapter, all statistics about mobile business shows that figures are improving, capacity is improving, and the environment for the development of mobile business is still improving in China. Thus, it is easy to generate from this chapter that China has been adopting every leading mobile technology all over the world, such as GPRS, CDMA, 3G, and Bluetooth, and not only adopting the advanced technologies; China is trying to put those mobile technologies into practice to produce productivity in many industry sectors, such as Entertainment, Education, Government Service, and so on. It is clear that mobile technologies benefit people throughout the world, including the people of China; we believe that the future of mobile business in China will only get better and better. We will continue to create more technologies and wait for the day when it blossoms out.

REFERENCES

- Beal, A., Beck, J. C., Keating, S. T., Lynch, P. D., Tu, L., Wade, M., & Wilson, J. (2001, June 4). *The future of wireless: Different than you think, bolder than you imagine*. Retrieved from http://www.accenture.com/xd/xd.asp?it=enWeb&xd=_isc/iscresearchreportabstract_134.xml
- Bing, W. (2004). Mobile government: The trend of e-government. *Nanfang Daily Newspaper*, (November 17), A-13.

- Blogchina. (2004). Retrieved October 5, 2004, from <http://www.blogchina.com/>
- Chen, L., & Nath, R. (2004). A framework for mobile business applications. *International Journal of Mobile Communications*, 2(4), 368-381.
- China Mobile. (n.d.). Retrieved October 5, 2004, from <http://www.chinamobile.com/>
- Choices. (n.d.). Retrieved October 15, 2004, from <http://choices.cs.uiuc.edu/MobilSec/>
- Christensen, G. E., & Methlie, L. B. (2003). Value creation in e-business: Exploring the impacts of Internet-enabled business conduct. *Proceedings of the 16th Bled E-Commerce Conference: E-Transformation*, Bled, Slovenia.
- CLi4 Report. (2004). Retrieved October 5, 2004, from <http://www.cl4.net/>
- CNNIC. (2003). *13th statistical survey on the Internet development in China*. Retrieved from <http://www.cnnic.com.cn/download/manual/en-reports/13.pdf>
- CNNIC. (2004). *Report*. Retrieved October 5, 2004, from <http://www.cnnic.com>
- Department of Informatization Promotion with the MII. (2003). *E-business development in China* (in Chinese). DIP Information Center.
- Feiwen, H. (2001, September 29). Good news about the PHS local wireless service by China Telecom (in Chinese). *Communication Information News*.
- Ghosh, A., & Swaminatha, T. (2001). Software security and privacy risk in mobile e-commerce. *Communications of the ACM*, 44(2).
- Guangzuo, C. (2004). *Mobile education: A new way for modern education technique*. Chongqing, China: Southeast Normal University Publishing House.
- Guangzuo, C., Doyong, S., Zhangliang, Zhangbo, & Zhangbaoli. (2002a). Simulation design for wireless mobile home robot based on LDPM. *Proceedings of the Korea-China Workshop of FIRA2002*, Korea.
- Guangzuo, C., Ypngsheng, H., Zhangliang, Dongyong, S., & Hu, C. (2002b). A case study on home robot based on Mobile Internet. *Proceedings of FIRA2002*, Korea.
- Huang, C. (1999, May). An analysis of CDMA 3G wireless communications standards. *Proceedings of IEEE Vehicular Technology Conference* (pp. 342-345), Houston, TX.
- Lange, D. B., & Serven, O. M. (1999). Good reasons for mobile agents. *Communications of the ACM*.
- MII. (n.d.). Retrieved November 20, 2004, from <http://www.mii.gov.cn/>
- Mobile Commerce World. (n.d.). Retrieved February 5, 2005, from <http://www.mobilecommerceworld.com/>
- Pahlavan, K., & Krishnamurthy, P. (2002). *Principles of wireless networks*. Englewood Cliffs, NJ: Prentice Hall.
- Rao, A. S., & Georgeff, M. P. (1995). BDI agents: From theory to practice. *Proceedings of the 1st International Conference on Multi-Agent Systems (ICMAS-95)*. San Francisco: ACM Press.
- SETC. (n.d.). Retrieved February 12, 2005, from <http://www.setc.gov.cn/>
- Shu, W. (2003). Wireless LAN and its application in internal and international education. *Modern Educational Technology*, 13(5), 18-21.

Stats. (n.d.). Retrieved February 12, 2005, from <http://www.stats.gov.cn/>

TechTarget. (2005). Retrieved from www.techtarget.com

Walke, B. H. (2002). *Mobile radio networks: Networking, protocols and traffic performance* (2nd ed.). New York: John Wiley & Sons.

Xiangdong, W. (2001). *Report on mobile communication & Mobile Internet in China*. Retrieved from http://www.telecomvisions.com/articles/pdf/china_mobile_internet.pdf/

Xiangliang, H. (2002). *Mobile education, a great vision* (in Chinese). Retrieved February 17, 2002, from <http://www.ccw.com>

Yuanrong, Y. (2002, November). An adaptability e-learning system based on Web. *Proceedings of the 8th Joint International Computer Conference*. Zhejiang, China: Zhejiang University Press.

Zmijewska, A., Lawrence, E., & Steele, R. (2004). Classifying m-payments—A user-centric model. *Proceedings of the 3rd International Conference on Mobile Business* (ICMB), New York.

Section VII

Application

Chapter XXXV

Developing Smart Client Mobile Applications

Jason Gan

University of Technology, Australia

ABSTRACT

This chapter examines the convergence of mobile technologies based on smart client architecture. To improve the usability and accessibility of mobile applications and services, the smart client architecture extends the capabilities of the mobile computing platform with support for multimodal interfaces, smart client database and synchronization, presence awareness, location awareness and identity management. Its broad impact on business communication and productivity is highlighted as a tangible benefit.

INTRODUCTION

In the highly competitive mobile market, a key differentiator is provided by improving the user experience of the mobile application. To improve the user experience, common usability and accessibility problems in mobile applications can be mitigated by multimodal interfaces and smart client architecture. For instance, the provision of multimodal interfaces for browser-

based applications can help to overcome the limitations of small viewing areas and input options, whereas smart clients based on rich application interfaces can be utilized to push processing load onto the mobile device. Furthermore, smart clients that enable presence, context sensitivity, location awareness, and real-time collaboration promise a new paradigm for mobile communications, delivering far richer, dynamic user experiences.

SMART CLIENT

As mobile enabling technologies advance in capability, affordability, and availability, users expect improved design of mobile devices that will leverage the advances and convergence in technology and the Internet to deliver richer applications and value-added mobile services (a.k.a., m-services). A key enabling technology for delivering on the promise of mobile applications with high levels of functionality, performance, flexibility, and integration is the *smart client*. This is a type of application model that bridges the gap between the thick and thin client models, providing the responsiveness and adaptability of a thick client model with the manageability of a thin client.

Dave Hill, from the Microsoft .NET Enterprise Architecture team, defines five characteristics of a smart client application (2004):

1. **Utilizes Local Resources:** Smart clients exploit local resources such as hardware for storage, processing, or data capture to deliver a richer user experience.
2. **Connected:** Smart clients are ready to connect and exchange data with various systems across the enterprise.
3. **Off-Line Capable:** Off-line capability using local caching and processing enable operation during periods of disconnection or intermittent network connectivity. Smart clients can send data in the background, resulting in greater responsiveness in the user interface.
4. **Intelligent Install and Update:** The smart client interface allows the remote update of the smartphone software to repair bugs, change characteristics, or incorporate new features.
5. **Client Device Flexibility:** Smart client applications support multiple versions that target specific device type and functionality.

The smart client architecture supports multimodality, data integration, Bluetooth interoperability, presence awareness, location awareness, and identity management. Each of these features extends serviceable functionality to the mobile application, from voice-activated commands to authentication and non-repudiation services. Moreover, the integration of serviceable functionality promises to deliver rich user experiences. For example, the voice-activated smart wireless device will automatically connect, authenticate, and show the identity and location of the receiver.

The convergence of presence, location, and identity management is an emergent technology integrating the services that support a secure mobile network and an online community environment with applications that facilitate information retrieval, communications, dating, gambling, financial management, trading, paying bills, games, and entertainment. As the mobile market is highly competitive and dynamic, and driven by the mass market demand for high-performance m-applications, the impact of technology convergence highlights the need for common industry standards.

INDUSTRY STANDARDS FOR SMART CLIENT MOBILE APPLICATIONS

As the specifications for smart client mobile application interfaces are complex, it is necessary for wireless developers to adhere to industry standards. The Mobile Industry Processor Interface (MIPI) Alliance is a non-profit organization that spearheaded the initiative of industry specifications for smartphones and application-rich mobile devices.

Wireless and embedded software developers have a choice between Microsoft .NET and Sun Java development frameworks and runtime

environments for designing and delivering next-generation mobile applications.

The Microsoft .NET Compact Framework is a subset of the developer software for PCs and servers that allows powerful .NET applications to run on handheld computers, and specifically supports: Pocket PC, embedded solutions running on Windows CE .NET for smart mobile devices, and Microsoft Smartphone 2002. The .NET Compact Framework can be extended to support additional mobile device interfaces. The inclusion of SQL Server CE 2.0 provides developers with a powerful, local relational database for creating dynamic, client-side mobile applications with database replication that enables remote devices to edit data in parallel. SQL Server 2000 supports two methods of replicating information from a back-end master to a remote client database: Remote Data Access (RDA) and merge replication (Thews, 2003).

The Sun Java platform includes the Micro Edition specification for building rich and smart client applications for mobile and embedded devices.

SMART CLIENT AND MULTIMODALITY

Smart client mobile applications equipped with multimodality and speech capabilities represent the next generation in portable office communications and wearable computer technology. Multimodality is defined as the optional presentation of the same information content in more than one sensory mode (European Communications Standards Institute, 2003). The multimodal interface on a multiple context-aware device enables interaction through different communication channels (a.k.a., modalities) to overcome the inherent input/output limitations of mobile devices. For example, a user

might use voice-only commands rather than the standard keypad input, while audio output can deliver additional information that cannot fit on the screen. The multimodal information is processed by the interpreter component of a multimodal host server.

Multimodal integration is the combination of different modalities to form a flexible user interface. There are seven types of modalities: visual, auditory, tactile, olfactory, gustatory, vestibular, and proprioception. Of these, visual and auditory are the most commonly used in m-applications. However, users who are blind and/or mute stand to benefit from the availability of tactile feedback and gesture modalities that address their physical disabilities.

SMART CLIENT DATABASE AND SYNCHRONIZATION

Reliable, secure, and immediate access to enterprise data irrespective of time and place is paramount in a time-critical business application, and it is especially critical in completing an online session—for example, when you have to close a bid on time. The problem with wireless technology is that data access and connectivity are intermittently connected and unreliable due to dropped connections, coverage issues, low bandwidth, and high latency. Smart client off-line access to data using a localized mobile database and synchronization with a central database (a store-and-forward mechanism) can help reduce and eliminate the performance bottlenecks related to slow and unreliable networks, and thus improve the user experience.

Besides improving the user experience, mobile applications can also affect the work environment. In particular, the collaborations of disparate teams stand to benefit from presence awareness.

PRESENCE AWARENESS

A study on disparate software project teams discussed the potential impact on productivity and the bottom line from leveraging presence and collaborative technologies in multi-site environments (Herbsleb, Mockus, Finholt, & Grinter, 2001, p. 9).

Presence is defined as a collection of real-time data describing the ability and willingness of a user to communicate across specific media and devices (Schneyderman, 2004). Presence awareness, a vital property of instant messaging applications, allows users to know when other users in a community are online and willing to exchange messages, what devices can be used for communications, and the real-time status of these devices. This can result in time and cost savings and improved productivity in many enterprise environments. For example, customer service environments stand to benefit from decreased operational costs and the ability to connect knowledge experts in real time across geographical locations and time zones. The adoption of presence technology into the workplace enables disparate workers to connect and communicate more efficiently by overcoming the lack of context and absence of informal communication (Herbsleb, Atkins, Boyer, Handel, & Finholt, 2002). A presence-based publish-and-subscribe channel can be deployed for discovery of available managers in workflow collaborations and for push content delivery.

It is important for developers to address the security and privacy concerns regarding presence technology. Building security and privacy controls into the presence-enabled device allows such features as access control, visibility, message blocking, and message encryption. Privacy policies restrict communication channels and prevent unsolicited conversations or uninvited listening. Biometric scanners built

into the presence-equipped device help to prevent such exploits as identity masquerading and spoofing. In addition, each receiving device would be configured to transmit its location and identity information to the mobile network.

LOCATION AWARENESS AND PERVASIVE COMPUTING

Location awareness in smart mobile applications is provided by mobile positioning technology such as the Global Positioning System (GPS). The GPS is a satellite navigation system that allows a mobile device with a GPS receiver to pinpoint a location on Earth by measuring the distances from a number of satellites simultaneously. Integration with GPS technology provides a mobile device with a pervasive-computing interface to location-based services (LBSs) such as emergency assistance and personal navigation. Assisted GPS (AGPS) describes a mobile positioning system that consists of the integrated GPS receiver and network resources such as an assistance server and reference network. The assistance server communicates with the GPS receiver via the cellular link and accesses data from the reference network, resulting in boosted performance of the receiver.

IDENTITY MANAGEMENT

Smart client mobile applications may require that access privileges to specific resources are granted only to users who have the right to use those resources. In an e-business system, access to certain resources is restricted to those who can supply the proper identity credentials by a process called authentication. Identity management is the federation of trusted endpoints that are able to authenticate each other's

presence at these endpoints with the intention of having a secure transaction (Smith, 2002). The banking and finance industry is a strong driver for this technology to reduce online fraud. Moreover, businesses and customers stand to benefit from expedited billing and payments. As identity and trust form the basis of successful e- and m-business, there is a strong demand from business and government sectors for efficient mobile integration of identity management services that will provide the end user with a more satisfactory online experience, enriched with higher levels of personalization, security, and control over identity information.

CONCLUSION AND FUTURE DIRECTION

The smart client provides an application model for developing richer, more responsive, and more usable mobile applications. Key areas of technological development based on the smart client are multimodality, presence awareness, location awareness, and identity management. These technologies describe a future to be defined by voice-activated systems, real-time collaboration, location-based services for information or emergencies, and mobile integration of identity services. Developing smart mobile applications that leverage the advances and convergence in technology is a step closer to realizing the full potential of a wireless Internet.

REFERENCES

- Antonopoulos, A. M. (2004, May 10). Location and presence take identity management to the next level. *Network World Data Center Newsletter*. Retrieved December 28, 2004, from <http://www.nwfusion.com/newsletters/datacenter/2004/1004datacenter1.html>
- European Communications Standards Institute. (2003). Human factors (HF); multimodal interaction, communication and navigation guidelines. *ETSI EG 202 191, 1.1.1* (August).
- Herbsleb, J. D., Atkins, D. L., Boyer, D. G., Handel, M., & Finholt, T. A. (2002). Introducing instant messaging and chat in the workplace. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Changing our World, Changing Ourselves*. Minneapolis, MN. New York: ACM Press.
- Herbsleb, J. D., Mockus, A., Finholt, T. A., & Grinter, R. E. (2001, May 12-19). An empirical study of global software development. *Proceedings of the 23rd International Conference on Software Engineering (ICSE'01)*, Toronto, Canada. Retrieved from [http://www-2.cs.cmu.edu/~jdh/collaboratory/research_papers/ICSE_01_final\(2\).pdf](http://www-2.cs.cmu.edu/~jdh/collaboratory/research_papers/ICSE_01_final(2).pdf)
- Hill, D. (2004). *What is a smart client anyway?* Retrieved December 12, 2004, from <http://weblogs.asp.net/dphill/articles/66300.aspx>
- Schneyderman, A. (2004). *Presence in mobile VoIP networks*. Retrieved from <http://www.tmcnet.com/voip/0904/featureshneyderman.htm>
- Smith, R. (2002). *Identity management—give me liberty or give me passport?* Retrieved December 28, 2004, from http://www.giac.org/practical/GSEC/Robert_Smith_GSEC.pdf
- Thews, D. (2003, October). Create mobile database apps. *Visual Studio Magazine*. Retrieved January 23, 2005, from http://www.ftponline.com/vsm/2003_10/magazine/features/thews/

Chapter XXXVI

Ontology-Based Information Retrieval Under a Mobile Business Environment

Sheng-Wei Guan
Brunel University, UK

ABSTRACT

The proposed OntoQuery system in the m-commerce agent framework investigates new methodologies for efficient query formation for product databases. It also forms new methodologies for effective information retrieval. The query formation approach implemented takes advantage of the tree pathway structure in ontology, as well as keywords, to form queries visually and efficiently. The proposed information retrieval system uses genetic algorithms, and is computationally more effective than iterative methods such as relevance feedback. Synonyms are used to mutate earlier queries. Mutation is used together with query optimization techniques like query restructuring by logical terms and numerical constraints replacement. The fitness function of the genetic algorithm is defined by three elements: (1) number of documents retrieved, (2) quality of documents, and (3) correlation of queries. The number and quality of documents retrieved give the basic strength of a mutated query, while query correlation accounts for mutated query ambiguities.

INTRODUCTION

Mobile computing will be the next buzzword of the twenty-first century. Presently, consumers demand personalized wireless computing ser-

vices while they are mobile. This infantile paradigm of mobile computing is opening up new markets. Corporate power users who are at the cutting edge of technology are always armed with an arsenal of mobile equipment.

CURRENT SITUATION AND MOTIVATION OF RESEARCH

According to Reuters and NUA Internet surveys in 1997 (Wieerhold, Stefan, Sergey, Prasenjit, Yuhui, Sichun et al., 2000), about 1.81% people worldwide surf the Internet for information daily. In the same year, according to surveys done by the Forrester Research (Wieerhold et al., 2000) and the Yankee Group, there was a significant increase of online retail sales, from \$600 million in 1996 to more than \$2 billion in 1997. This value reached \$282 billion in 2000 and is still increasing. With the exponentially growing number of Internet users over these few years, the International Data Corporation (IDC) expected an increase to \$4.3 trillion by 2005. Thus, as can be seen, trading online has become increasingly important to the commercial world. It is inevitable that e-commerce will be the next strategy that companies will adopt.

At the same time, with the introduction of new technologies such as WAP, HSCSD, GPRS, UMTS, and Bluetooth, together with new and personalized applications, it is believed that the e-commerce arena will sooner or later merge its applications with handheld devices to create more opportunities for the birth of mobile commerce. In fact, research from the IDC expected the mobile portal to reach 55 million users by 2005.

However, according to the IDC, there is a 26% drop in the sales of handheld devices in the first quarter of 2002. One of the reasons why the potential of mobile commerce is largely unrealized to date is because a single killer application that can attract wireless users to use wireless services still does not exist. According to a recent survey by Gartner, Inc., besides the importance of coverage of wireless network and pricing issues, the wireless Internet and data services is the next crucial factor that attracts users to use wireless service. As such,

there is a need to improve data services over the wireless network. One of these services is the information retrieval service.

Most electronic product information retrieval systems are still not efficient enough to cater to the increasing needs of customers. Typically, as product information in the Web soared eminently, reusing and sharing of product information has become extremely important. This is especially true in the m-commerce arena where the bandwidth of mobile devices is low and large data would not be possible. Thus, the discovery of new information retrieval techniques is inevitable. Also, observations and studies have shown that the average user often selects inappropriate information retrieval resources and uses them inefficiently and ineffectively. People seem to be content to retrieve any information on their topics, regardless of quality. Few people currently recognize the need for improving their information retrieval skills. Hence, there is a need to simplify the way people form queries to retrieve information.

OBJECTIVES AND RESEARCH CONTRIBUTION

The main objective of this chapter is to improve information retrieval services for the m-commerce arena. After considering the flaws in current information retrieval systems, this chapter proposes a methodology for efficient query formation for product databases in m-commerce. In addition, this chapter proposes a methodology for effective information retrieval systems, which includes the evaluation of retrieved documents to enhance the quality of results that are obtained from product searches.

This chapter discusses the usage of ontology to create an efficient environment for m-commerce users to form queries. The establishment of a method that combines keyword searches while using ontology to perform query

formation tasks further allows a more flexible m-commerce environment for users. Also, with the use of genetic algorithm, it is hoped that query effectiveness can be achieved, at the same time saving computational time. In this chapter, new ways of defining the fitness function of the genetic algorithm are also explored.

DEFINITION OF ONTOLOGY

In philosophy, ontology is a theory about the nature of existence, of what types of things exist; ontology as a discipline studies such theories. In the world of artificial intelligence, ontology (Fensel, 2000; Braga, Werner, & Mattosso, 2000; Hendler, 2001) is defined as a design of a conceptualization to be reused across multiple applications. A conceptualization is a set of concepts, relations, objects, and constraints that define a semantic model of some domain of interest. In other words, ontology is like the structure that describes or encodes the conceptualization in any relational aspect.

QUERY FORMATIONS

Query Formation Using Keywords

The traditional way of searching is to use keywords as a guide to form and process queries. Query formation using keywords is just fine if the user knows exactly what he or she is looking for. However, it may be difficult or impossible to use keyword-based search when the vocabulary of the subject field is totally unfamiliar to the user. Furthermore, using the same keywords in a different order or at a different time might change the number as well as the order of sites listed. As can be seen, there is usually some form of uncertainty when forming a query using keywords. In addition,

because a word can contain different meanings, ambiguities will arise due to missing context. Moreover, no single engine searches the entire Internet, so it is often necessary to search several engines. Essentially, the element that is lacking in keyword-based queries is the user's sense of linkage between multiple keywords and context. Also, if too many keywords are used, the constraints may become too tight to retrieve anything.

Query Formation Based on Agent Ontology in Bioinformatics

A relatively new approach is to form a query using ontology where ontology serves as a context and structuring mechanism for keywords. Functional bioinformatics is an emerging sub-field of bioinformatics that is concerned with ontology and algorithms for computing with biological functions (Karp, 2000). It uses regulatory pathways to define ontology structures. In graph theory, a regulatory pathway is usually represented as an oriented graph with vertices corresponding to substances and edges corresponding to interactions (Rzhetsky, Tomohiro, Sergey, Shawn, Michael, Sabrina, et al. et al., n.d.). This is similar to what we find in tree ontology—the nodes represent the substances and the links between nodes represent the interactions.

Functional bioinformatics focus on large-scale computational problems such as problems involving complete metabolic networks and genetic networks (Karp, 2000). This allows information retrieval algorithms to evaluate queries easily (McGuinness, 1998). Also, functional bioinformatics is concerned with representing, visualizing, and computing functional descriptions of individual genes and gene products (Karp, 2000). Having such an archive with organized ontological structures, ontology can be used to design knowledge portals for manual browsing (McGuinness, 1998). Thus, any user

can easily generate a query by just a few mouse clicks on the displayed ontology and then submit it to the database to get required information. Many Web catalogs, such as Yahoo, employ ontology in order to organize their contents (McGuinness, 1998). This type of query formulation would not be possible without semantic representation of the ontology. Ontology often plays the role of controlled indexing languages. Moreover they may be used in assisting the query formulation process, or for expanding queries with synonyms, hyponyms, and related terms in order to improve recall (McGuinness, 1998). Ontology can grow and shrink as necessary based on the context. In different contexts, parts of an ontology can be hidden or another made visible so that a new view of the same information space can be generated to suit a certain group of audience. A variety of smaller ontologies can be created by experts from various fields independently of each other, and then later can be merged with little effort to create a bigger ontology that can serve a different or larger audience.

INFORMATION RETRIEVAL

Relevance Feedback

Relevance feedback (Salton, 1989) is still the main technique for query modification. This technique has been investigated for more than 20 years in various information retrieval models, such as the probabilistic model (Robertson & Sparck Jones, 1976; Robertson & Walker, 1997; Haines & Croft, 1993) and vector space model (Salton, 1970; Boughanem, Chrisment, & Tamine, 1999; Salton, 1989). Relevance feedback is based on randomly changing the set of query terms as well as the weights associated with these terms according to the document retrieved and judged during the initial search. Conventionally, retrieved documents

are not evaluated. This means that a user might end up getting junk instead of the things he wants. Also, when there are too many documents available, the user has to filter them manually. These problems are fundamental to the motivation of this research.

Recently, many scientists were gathering a genetic approach to solve the information retrieval problem. However, there is still a great deal of room for research on genetic algorithms as there are many ways the fitness function can be defined.

Genetic Algorithm

A lot of research has been done on how genetic algorithm can be used in information retrieval. One popular approach is the restructuring of queries. Restructuring queries is necessary to improve the efficiency and effectiveness of the queries formed. This boils down to the fundamental concepts in relevance feedback. Hence, genetic algorithm actually extends the concepts of relevance feedback. The difference is that genetic algorithm uses more than one query and compares the fitness among these queries. The fittest query will survive in the end. This is much more effective than when only one query is used where hill-climbing search is used (Boughanem et al., 1999). Here, mutation and crossover take place during restructuring.

Yang and Korfhage (1994) use genetic algorithms to search the space of possible queries to generate better queries based on relevance feedback for the vector space model, and they obtained remarkable results. Kraft (Salton, 1989) explored to extend the concepts to deal with Boolean queries. Another approach uses evolutionary algorithms to evolve agents instead of queries (Kouichi, Toshihiro, & Hiroshi, 1999). These agents contain search parameters that will search all Web databases and are tuned up by genetic algorithms.

In this chapter, the focus is to extend the concepts of using genetic algorithms in query restructuring. The difference lies in the type of restructures. Basically, three types of restructures—using synonyms, logical operators, and numerical constraints—are used in the search process to enhance query effectiveness. Each type of restructure is rendered under different situations depending on the suitability in that situation.

The Fitness Function

Salton and McGill (1983) consider that the two ultimate measures of query fitness are namely precision and recall. Precision is the percentage of documents retrieved that are relevant, while recall measures the percentage of the relevant documents retrieved (Kraft, Petry, Buckles, & Sadasivan, 1994; Salton & McGill, 1983). These two tend to be inversely proportional so that one is traded for one another in most situations. They are complementary and competitive (Kraft et al., 1994). A more general query retrieves more documents, and recall increases because of more relevant documents retrieved. As the query becomes less general, fewer documents are retrieved and thus precision increases. Therefore, usually these two measures are not used directly. In Kraft et al. (1994), two arbitrary constant weights were used to balance the fitness function.

Another measure is the average search length (Losee, 1991). The average search length is the average number of documents or text fragments examined in moving down a ranked list of documents until arriving at the average position of a relevant document (Losee, 1988, 2000). Evaluating the performance of a filtering or retrieval process with the average search length provides a single number measure of performance. Also, it is capable of being predicted analytically (Losee, 2000) and is easily

understood by a system's end users. This is unlike most of the other retrieval and filtering measures based on precision and recall that are used to evaluate retrieval systems.

Another fitness function is the average maximum parse length (Losee, 2000). The average maximum parse length is the average (over a set of sentences) of the largest number of terms in a parse for each sentence. There are also measures that combine both the average search length and the average maximum parse length. The motivation for it is that having different genes producing the same average search length value is very common, and adding the average maximum parse length breaks the ties that exist when using average search length alone as the fitness function (Losee, 2000). When two genes with identical average search lengths are compared, the gene with the larger average maximum parse length is selected. This computational method is useful when the average search length is not a very good measure of the quality of parsing performance due to the limited improvements obtained with disambiguation. In Losee (2000), a useful fitness measure was obtained by weighting the two values such that that one-hundredth of the average maximum parse length was added to the negation of the average search length. The greater the value for this combined function, the fitter the gene.

In review to the present ways of defining the fitness function for a given genetic algorithm, the relevance of the documents retrieved has been greatly emphasized. Typically, present methods only dealt with the relevance of the document retrieved. This is reasonable but inefficient, because it is rather difficult to indicate the relevance of a document when the number of documents could be very large. In fact, relevancy can also exist at the query stage. This chapter measures the relevance of queries instead of documents retrieved. When this happens, the efficiency will improve signifi-

cantly as the number of queries will be much smaller than the number of documents retrieved. Of course, it might be true that a relevant query might not retrieve a relevant document. However, in the context of this chapter, where the retrieved documents are products from product databases, it is reasonable to say that almost all the documents retrieved are relevant to the query formed. This is because most product databases have rather fixed formats, unlike documents retrieved from the Internet where the variations of formats can be infinite. In addition, when a query is formed from a product ontology, it is obvious that the linkage between the terms has rid off much of the irrelevancy of the documents retrieved. Thus, the fitness functions used in the present methods are not used in this chapter.

THE PROPOSED APPROACHES

Combining Keyword Queries with Ontology

Both keyword- and ontology-based approaches have their advantages and disadvantages. Ontology provides the structure, context, and visual aid, while keyword provides a direct search mechanism. Thus, by combining keyword-based queries with ontology, it is possible to achieve a better and more effective query formation. Basically, most would think of a parallel combination that means the keyword approach can be accessed as a backup method to the ontology approach. For example, when a user cannot find a term in the ontology, he may still be able to use keywords to search whatever is not covered by the ontology.

The parallel solution can be considered in terms of its coverage, but still lacks efficiency in some ways. For example, as the ontology gets larger, finding a term becomes a chore and slows down everything. To tackle this problem,

a serial combination will be needed. This is the part where most research contents can be found, as this method—although similar—does differ from the most current research.

On the whole, before the ontology terms are accessed to form the queries, there will be a keyword search to find the required ontology term. For example, “ps2” can be hidden in the node “mouse” when presented in the ontology. The user will not be able to know where “ps2” can be found intuitively without eyeballing the ontology. With the help of keyword search, the term “ps2” can be found easily.

Query Restructuring

In forming queries, there can be a great chance that the vocabulary used by the user to describe a query does not exactly match the vocabulary used by a query system. This will result in getting insufficient information. For example, a person looking for televisions might key in “television” for his search, but the product ontology might describe televisions in a node called “tv” or “tv set.” Intuitively, “television” is equal to “tv” and “tv set.” Therefore, restructuring dealing with domain ontology relationships might be useful. These relationships involve semantic links such as hyponyms and synonyms (Braga et al., 2000). Here, using synonyms is an adequate option to restructure queries.

Also, one major problem about information retrieval is that either too little or too much information is retrieved when making a query. When too little information is retrieved, it is logical to relax the constraints of the query. Thus the use of synonym or hyponym might be necessary. However, this approach has a major disadvantage. By relaxing the constraints of a query using synonym or hyponym to increase the number of documents retrieved, one could actually deface the meaning of the original query such that it could drift away from the

user's intention. This concern can be alleviated by having user feedback along the process; there is still much to research on the use of the synonym or hyponym. Another new approach that we have considered in this research is to relax constraints step by step. This option can better eliminate the chances of constructing far-fetched queries from the use of genetic algorithm. For example, step by step means to relax numerical constraints in steps of 5% or 10%. For instance, $\text{price} < 2000$ becomes $\text{price} < 2200$. In addition, instead of relaxing constraints, this approach can also tighten constraints. The only limitation that this option has is that it is only applicable to numerical values. When the constraints are words, this method might not work.

Of course, one can mutate wordy constraints by removing or truncating a few letters or even words from them. But, the same problem about ambiguities will set in.

The Fitness Function

As mentioned earlier, since the relevancy focus has been shifted to the query stage, and product documents are mostly relevant when the query is relevant, it is not necessary to include measurements like precision or recall, which measure relevance at the document-retrieved stage. In this chapter, a totally new design of the fitness function is schemed.

Instead of favoring a query by giving it credits for relevancy, it is a good idea to work from the irrelevancy of the query by giving it demerits since most documents are relevant. Here, the irrelevancy of the query will set in after the mutation of the queries is applied when using genetic algorithm. This irrelevancy will form one of the elements in the fitness function.

Also, addressing the issues of finding too many or too few documents, it is functional to include a weight element that measures the

number of product documents that a query can obtain in its fitness. Another weight element is the quality of the product documents that a query can retrieve. This element should be included simply because it is rather easy to measure the quality of product documents than to measure other documents from the Internet because of some numerical values. In making queries on products, it is common to encounter numerical constraints where the values are measurable. Of course, this element is limited such that it can only measure the fitness of queries with numerical constraints. The details of the complete design of this fitness function are discussed later in this chapter.

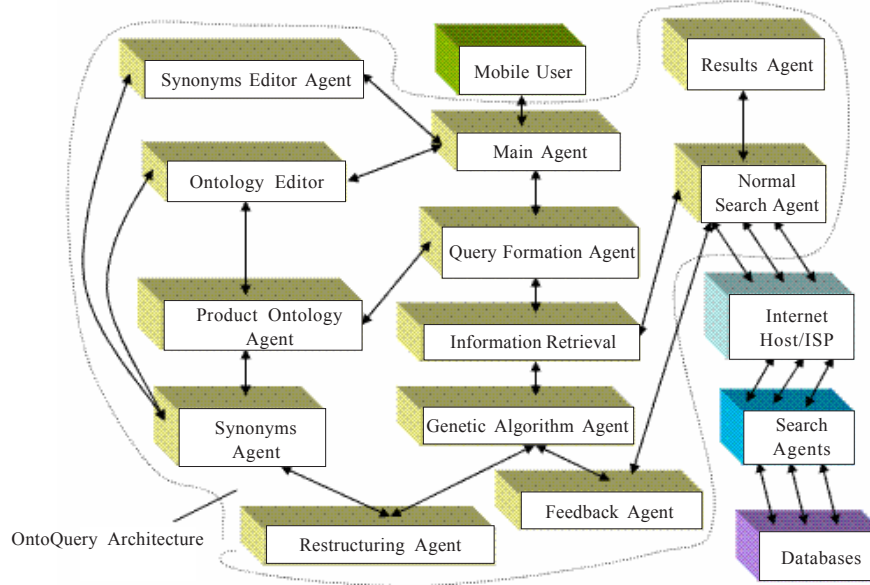
PROTOTYPE DESIGN AND IMPLEMENTATION

Overall Architecture

Basically, the architecture of the research problem lies largely on the design of the agents. It is important to know exactly what the agents are that will be needed before drilling into the problem. Figure 1 shows a system architecture of the relationships among the modular agents.

Each mobile user can activate the Main Agent of the OntoQuery system. On the whole, three major application agents—the Synonyms Editor Agent, the Ontology Editor Agent and the Query Formation Agent—are interfaced to the Main Agent. The Synonyms Editor Agent merely interacts with the Synonyms Agent by editing the semantics of the synonyms. The Ontology Editor Agent interfaces with the Product Ontology Agent to allow the user to edit his own product ontology. At the same time it saves the new ontology terms with the help of the Synonyms Agent to prepare the synonyms for editing. The Query Formation Agent uses the Product Ontology Agent to aid the user to form queries. It also calls for the Information

Figure 1. System architecture among agents



Retrieval Agent, which is interfaced directly to the Normal Search Agent and Genetic Algorithm Agent. The Genetic Algorithm Agent employs the Restructuring Agent to restructure query population according to the feedback and requirements received from the Feedback Agent. Meanwhile, the Restructuring Agent does query restructuring with the help of the Synonyms Agent. The Feedback Agent also links up the Normal Search Agent, which will submit queries to retrieve information wirelessly from the databases in the Internet through the Internet service provider. In addition, there will be Search Agents residing at the host or the Internet service provider, which provides the information retrieval service its connected databases. Here, many mobile users are connected to the Internet via the OntoQuery architecture, while the Internet connects to many databases. The Normal Search Agent will then display the retrieval results via the Results Agent.

Query Formation Using Ontology

Query formation is a rather straightforward application; there will be multiple requirements just like present query systems in the industry. The challenges here lie in the way queries are created and the format in which they will be

Figure 2. Illustration of using ontology to form queries

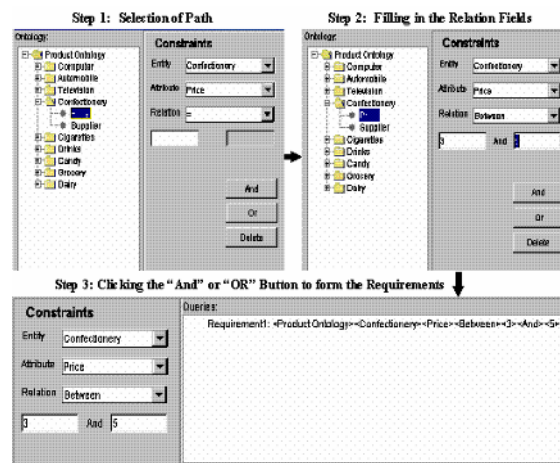
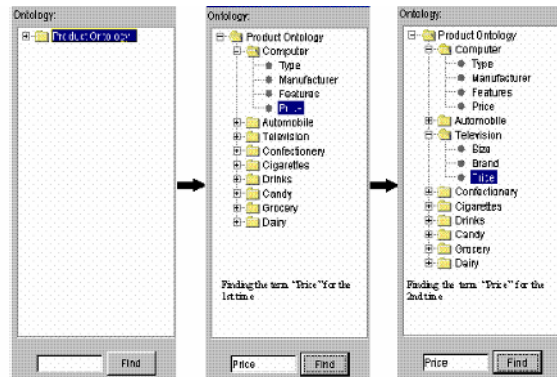


Figure 3. Illustration of the sequence of events for finding ontology terms



sent to the Search Agents. Here, query formation will be done with the aid of tree ontology. Multiple requirements help form a complete query. An illustration of the query formation process is shown in Figure 2.

As can be seen from this illustration, using ontology to form a query has become a simple, three-step procedure. This helps the user save several steps by forming a query using the ontology path that is selected. Thus, it can be claimed that forming queries using ontology is more efficient than using keywords.

Combining Keywords and Ontology

The design of parallel combination is rather straightforward. An ontology does not cover everything. Thus, besides having ontology for the user to click on when forming a query, there should be some fields present for the user to fill in. When these fields are being filled in, they can either replace the use of ontology either partially or completely.

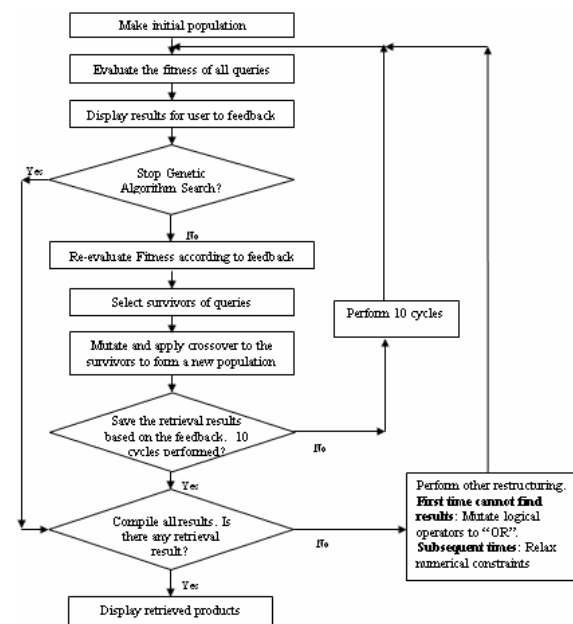
For a serial combination, keywords are used to look for ontology terms in the ontology. This is necessary because when the ontology is too large, search for an ontology term by manual

clicking becomes difficult. Thus, there would be a field that allows the user to highlight the terms in the ontology itself. This process is similar to how we search for words in Microsoft Word documents, except that now the search is to transverse in an ontology tree structure, as shown in Figure 3. From this illustration, it can be seen that using keywords to search ontology terms in the ontology creates an efficient environment and context for the user. It increases the chances for getting results from a tedious search of ontology terms, especially when the product ontology gets bigger.

Information Retrieval

Using the query formed by the query formation application, an application searches the databases to retrieve information. Intuitively, this application would first do a normal search before allowing the user to proceed with a genetic algorithm search. This is because a

Figure 4. Flowchart of the searching process system architecture among agents



genetic algorithm search would definitely take a much longer time than a normal search because of its expensive iterations. The retrieval results are presented to the user and if he is not satisfied, he can then choose to proceed with a genetic algorithm search.

Genetic Algorithm

If the user requests the use of a genetic algorithm, the system will request some input from the user to perform genetic algorithm computation. The system then creates a population of queries from the original query. Figure 4 shows the flowchart of the sequence of events.

Basically, genetic algorithm will mutate the queries according to the synonyms of the terms in the ontology. This is necessary in order to retrieve all the relevant documents that are related to the query. Thus, there will be a lookup table of synonyms. This creates semantics to relate the synonyms with the ontology terms. Each term is a gene, and each query forms a chromosome.

The Fitness Function

The main concern of using genetic algorithm is the design of the fitness function. In this application, three major elements are used to define the fitness function. They are the fitness of the number of documents retrieved (f_d), the fitness of the average quality of the query results (f_q), and the overall correlation for the query (f_r). The fitness of each chromosome is calculated as follows:

$$Fitness = |f_r(f_d + f_q)|$$

| . | indicates that the fitness function is normalized to form a population distribution function.

The calculation of the value of f_d is not as straightforward. Since, the drawback is getting

either too few or too many documents, the user should be able to specify the ideal number of documents (i). Typically, if the user does not know the value of i , the default value will be 20. Using the value of i as the mean value, two ‘band pass filters’-like functions, namely the triangular and Gaussian functions, are used to create a more flexible mapping from number of documents retrieved (d) to f_d . The triangular function gives a constant drop in “gain” (decrease in fitness) from the ‘center frequency’ (the mean). This function is good when the user wants to give an equal amount of demerits for every document that is away from his expected or ideal number. By specifying the value of the gradient for the right side of the triangle, the user can determine how many demerits to give for every document that exceeds his expected number. When the value of this gradient is 0, the triangular function will look like a “high pass filter”. The Gaussian function is a more robust or high-ordered “band pass” such that its “bandwidth” could be specified by the value of the standard deviation. Thus, this function is useful when the user wants to give heavy demerits to queries that do not fall near his expected or ideal number. Also, when the standard deviation is large, the function will look like a ‘low pass filter’.

Only requirements that are specified by numerical constraints will have the value f_q . For example, documents retrieved according to *price < 2000* have numerical results and would retrieve results under the columns of price while documents retrieved according to *brand = Sony* do not have numerical results under the columns of brand. Here, it is required that the numerical values are summed up and averaged. Then, they are normalized. In addition, there is another consideration. The signs “<” and “>” formed during the query indicate the direction towards which the quality is favored. For example, when encountered with a “<” sign, it is logical to give higher fitness to the

chromosomes with smaller values in the numerical results. Here, a reciprocal function ($1/x$) is used before normalization takes place.

The next interesting portion that contributes to the fitness function is the correlation of the synonyms (f_r) with the ontology terms. A value from 0 to 1 is assigned to each relation between the ontology terms and their synonyms. For example, the discrete correlation between the terms “Television” and “TV” can be 0.9, while the discrete correlation between the terms “Price” and “SP” can be 0.7. When a requirement is $\langle \text{Television} \rangle \langle \text{Price} \rangle \langle \langle \rangle \langle 2000 \rangle$, the value of f_r will be the product of all the discrete correlations. In this case, the value of f_r is 0.63. Also, the user should be able to edit the correlation values to his preference. In addition, when there are many requirements in the query, these requirements will be linked with an “OR” or “AND” term. In this case, multiplying these correlation values together would not be rational. This is because, when too many terms are different, the value of f_r would become very small. In order to alleviate this problem, using maximum and minimum approaches would be more suitable than the product approach. For example, when there are two requirements A and B with correlations f_{r1} and f_{r2} respectively, a query of A “OR” B would give an overall correlation of $\max(f_{r1}, f_{r2})$. Conversely, a query of A “AND” B would return an overall correlation of $\min(f_{r1}, f_{r2})$.

In a multi-objective requirement, the long term will be split as two or more terms whereby

the paths of the tree become longer. For example, $\langle \text{Alarm Clock} \rangle \langle \text{Selling Price} \rangle \langle \langle 5 \rangle$ is single objective while $\langle \text{Clock} \rangle \langle \text{Alarm} \rangle \langle \text{Price} \rangle \langle \text{Selling} \rangle \langle \langle 5 \rangle$ is multi-objective. In fact, the maximum and minimum approaches should be combined with the product approach if the requirement is multi-objective. This is because if we multiply the correlation of $\langle \text{Clock} \rangle \langle \text{Alarm} \rangle$ and $\langle \text{Timepiece} \rangle \langle \text{Ring} \rangle$ because both terms are different, the problem of small r -value will surface again. However, in this chapter, the requirements are single objective. Thus, this combination of approach is not employed.

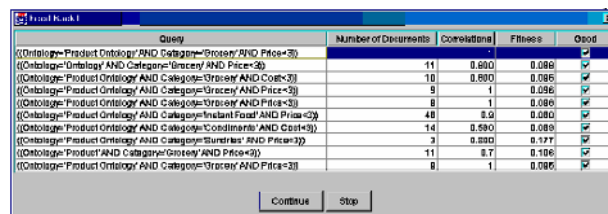
Mutation and Crossover

The concept of mutation is to replace some terms with synonyms when parsing the results. By doing so, there will be a lesser chance that some information is missed because the query will be more comprehensive. Basically the mutants are the terms that are included in each query. These terms are mutated randomly according to the synonyms so that new populations will be formed. Crossover will only be interchanging the different genes between two different chromosomes. A one-point crossover will be performed. This is also done randomly.

Feedback and Selection of Survival

The survivors are selected according to their overall fitness in the roulette-wheel selection

Figure 5. Screenshot of a feedback frame



Query	Number of Divariants	Correlation	Fitness	Good
(Ontology:Product Ontology AND Category:Grocery AND Price=30)	11	0.880	0.088	☒
(Ontology:Product Ontology AND Category:Grocery AND Price=30)	10	0.880	0.088	☒
(Ontology:Product Ontology AND Category:Grocery AND Price=30)	9	1	0.088	☒
(Ontology:Product Ontology AND Category:Grocery AND Price=30)	8	1	0.088	☒
(Ontology:Product Ontology AND Category:Instant Food AND Price=30)	48	0.0	0.080	☒
(Ontology:Product Ontology AND Category:Condiments AND Price=30)	14	0.880	0.088	☒
(Ontology:Product Ontology AND Category:Snacks AND Price=30)	3	0.880	0.127	☒
(Ontology:Product AND Category:Grocery AND Price=30)	11	0.7	0.106	☒
(Ontology:Product Ontology AND Category:Grocery AND Price=30)	8	1	0.088	☒

Other Restructures

Besides replacing query terms using synonyms, logical operators and numerical constraints can be restructured. This is necessary when the system cannot retrieve any document with its synonyms. However, it is still necessary to inform the user and consult his permission to carry on with this restructuring.

The system will mutate “AND” operators that link each requirement to “OR” operators. Logically, no documents will be retrieved if two far-fetched requirements are submitted as a query. Thus, mutating a logical “AND” term to an “OR” term can be a good choice to relax query constraints and yet it does not affect the originality of the query by much. In addition, the system can mutate numerical constraints. The system applies step-by-step relaxation to numerical constraints in a query until some documents are found. An illustration is shown in Figure 6.

This illustration is performed at steps of only 5%. As can be seen, because of the compounding effect, one item could be found during the fifth iteration of the process. Obviously, this method is still an effective method in terms of query restructuring because it is able to retrieve something which may be useful for the user.

Application of OntoQuery in M-Commerce

The screenshots used for illustrations are implemented using Java. One major consideration here is that mobile devices tend to have a much

smaller screen. Therefore, in order to realize the full OntoQuery architecture, there is a need to scale down some of the display size. Another possible solution is using scrollbars to view the screen, as shown in Figure 7.

PROTOTYPE TESTING AND EVALUATION

Effectiveness of the Genetic Algorithm

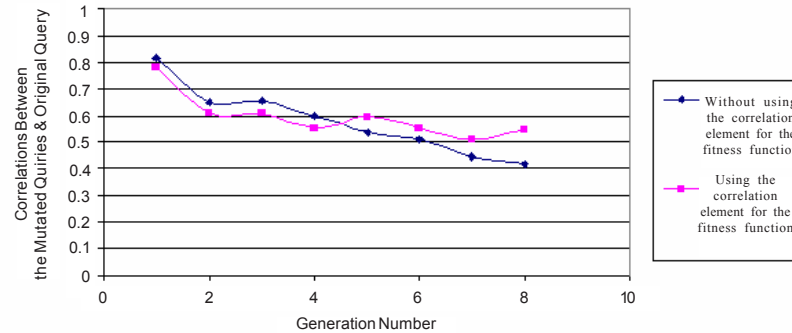
It is believed that the effectiveness of the genetic algorithm chosen is mainly determined by its supremacy in query effectiveness amplification. This is because its evolution power allows more retrieval results. The system was tested with a product list database. The effectiveness was measured by testing a series of queries with and without using genetic algorithm. For example, a query “<Product Ontology><Grocery><Price><<><3>” can only retrieve nine items when a normal search was performed, but can retrieve 98 items when genetic algorithm was performed. Table 1 shows the other results obtained by other queries.

By comparing the results shown in Table 1, it is obvious that using genetic algorithm does in fact retrieve more items than using a normal search. Thus, it is logical to say that using genetic algorithms will provide a more flexible and effective solution to information retrieval systems.

Table 1. Results showing effectiveness of GA

Query Formed	Without GA	With GA
<Product Ontology><Drinks><Price><<><2>	25	95
<Product Ontology><Diary><Price><<><3>	3	92
<Product Ontology><Candy><Price><<><3>	13	178
<Product Ontology><Confectionery><Price><<><3>	20	248
<Product Ontology><Confectionery><Supplier> <Contains><Ho>	1	52

Figure 8. Trends for the correlations between mutated queries and original query



Effects of the Fitness Function

The fitness function in a genetic algorithm determines how well it can optimize a query. The OntoQuery system tested out various fitness functions to improve the power of the genetic algorithm. The usage of triangle or Gaussian functions to evaluate the fitness for the number of documents retrieved suggested some ways to counter the “too many or too few retrieved documents” dilemma in typical search engines. The fitness for quality allows the assessment of queries based on what they can retrieve. This gives a good weight to prevent the fitness for the number of documents retrieved to dominate the fitness function.

It is thought that the introduction of using correlation prevents the original query from mutating into irrelevant queries. Implementing a choice to select whether to include the correlation fitness in the fitness function in OntoQuery creates a chance to test this claim. Figure 8 shows the graphs of the trends. From the trendlines, it can be seen that a converging trend was achieved such that the queries will still be quite relevant when correlation fitness is used. A more diverged or decreasing trend for the mutated queries was obtained when correlation fitness was not included. This proves that the use of correlation can prevent the original query from mutating into irrelevant queries.

Efficiency of the Genetic Algorithm

Although using genetic algorithm allows a more flexible and effective platform in retrieving information, there is no doubt that it trades off efficiency due to its expensive iterations. Thus, the only study that can be made here is about its improvement over relevance feedback. In relevance feedback, query expansion is achieved by modifying a query. Similarly, genetic algorithm extends the relevance feedback techniques with an addition rule, the survival of the fittest.

In this research, it is thought that several factors affect the efficiency of genetic algorithm. These anticipated factors include the population size used, the number of generations, and the length of the query. Here, the efficiency of the system is measured as follows:

$$\epsilon \text{ (Efficiency)} \approx \frac{E}{t} \approx \frac{D}{I}$$

where E denotes the effectiveness of the system
 t denotes the time taken for the system
 D denotes the number of relevant

Figure 9. Graph of efficiency vs. population size

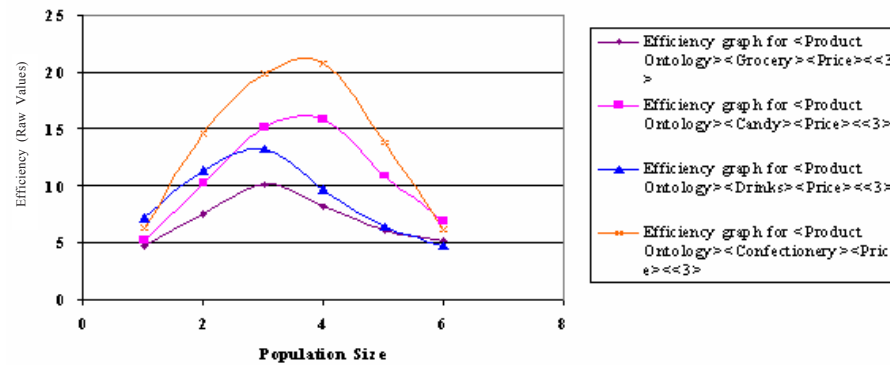
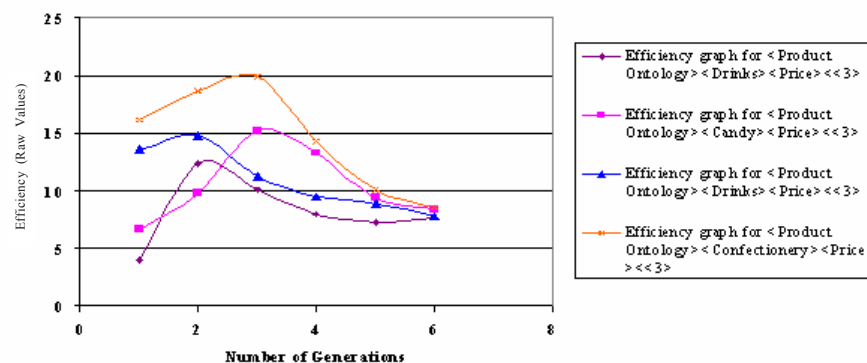


Figure 10. Graph of efficiency vs. number of generations



documents retrieved
 I denotes the number of iterations

Efficiency is formulated as above because it is believed that the number of documents retrieved is linearly proportional to the effectiveness of the system. Also, the number of iterations is directly related to the time taken to retrieve the results.

Effect of the Population Size

This test involved the comparison of the results obtained by varying population size from values 1 to 6. A few queries were used to find out the optimal population size by comparing the efficiency obtained when each population size was used. The number of generations is set to 3.

Figure 9 shows the results obtained by averaging three samples.

From Figure 9, it can be seen that the efficiency initially increases with the population size, but eventually decreases for all the queries used in this test. Every query has an optimal population size. However, these optimal values are not the same. In addition, the optimal efficiencies peaked at different values. A study was made on these findings, and it was found that the optimum value of the population size depends much on the uniformity of the database. For example, in the category “Confectionery” or “Candy”, the number of items that are contained under the category of each synonym in the database is more uniform than in the category of “Drinks” or “Grocery”. Also, the optimal efficiency value is higher when the

Figure 11. Effects of longer queries

Original Query (1 Requirement): ((Ontology='Product Ontology' AND Category='Grocery' AND Price<3))		
Population Size	3.0	Total Number of Documents 71 Efficiency 11.833333
Original Query (2 Requirements): ((Ontology='Product Ontology' AND Category='Grocery' AND Price<3)) OR ((Ontology='Product Ontology' AND Category='Drinks' AND Price<3))		
Population Size	3.0	Total Number of Documents 146 Efficiency 24.666666
Original Query (3 Requirements): ((Ontology='Product Ontology' AND Category='Grocery' AND Price<3)) OR ((Ontology='Product Ontology' AND Category='Drinks' AND Price<3)) OR ((Ontology='Product Ontology' AND Category='Candy' AND Price<3))		
Population Size	3.0	Total Number of Documents 183 Efficiency 30.5
Original Query (4 Requirements): ((Ontology='Product Ontology' AND Category='Grocery' AND Price<3)) OR ((Ontology='Product Ontology' AND Category='Drinks' AND Price<3)) OR ((Ontology='Product Ontology' AND Category='Candy' AND Price<3)) OR ((Ontology='Product Ontology' AND Category='Candy' AND Price<3))		
Population Size	3.0	Total Number of Documents 222 Efficiency 36.633332
Original Query (5 Requirements): ((Ontology='Product Ontology' AND Category='Grocery' AND Price<3)) OR ((Ontology='Product Ontology' AND Category='Drinks' AND Price<3)) OR ((Ontology='Product Ontology' AND Category='Candy' AND Price<3)) OR ((Ontology='Product Ontology' AND Category='Candy' AND Price<3)) OR ((Ontology='Product Ontology' AND Category='Candy' AND Price<3))		
Population Size	3.0	Total Number of Documents 284 Efficiency 47.333332

database is bigger. This is reasonable, as more results would be returned.

Effect of the Number of Generations

This test is similar to the test on the effect of population size except that here, the number of generations was varied from values 1 to 6. The population size of this test is set constant at a value of 3. Figure 10 shows the graph of the results obtained by averaging three samples. Again, with resemblance to that of the population size test, efficiency initially increases with the number of generations, but eventually decreases. The difference between the two tests lies in the optimal values. It is observed that the optimal numbers of generations are smaller than the optimal population sizes. Indeed, by comparing Figures 9 and 10, it can be observed that the efficiency for a larger number of generations is better than the efficiency for greater population sizes. The explanation for this is that if queries go through more generations of competition, the queries become fitter after each round and hence gives better results.

Effects of Longer Queries

Logically, a longer query would unleash more power in genetic algorithms. This is because there will be more synonyms qualified for query modification. In this test, queries with two or more requirements are tested. The population size is set to 3 and the number of generations is set to 2. The result is compared with the result with only 1 requirement. The screenshot of the results is shown in Figure 11. As can be seen from Figure 11, efficiency increases as the number of requirements increases. Thus, it is reasonable to say that our genetic algorithm is more suitable for longer queries to be efficient.

CONCLUSION AND FUTURE IMPROVEMENT

In summary, this research work investigated the OntoQuery system within an m-commerce agent framework against current query formation and information retrieval systems.

The prototype implementation results showed that querying formation using an ontology ap-

proach is efficient as it provides a friendly environment to the user. In addition, by combining the keyword and ontology approaches, a more efficient and effective way of forming queries could be achieved. Thus, the objective to propose efficient query formation for product databases is successful.

It was found that genetic algorithm is able to optimize queries effectively. Also, using genetic approaches, the OntoQuery system proposed and tested out various fitness functions for searching product databases. The use of correlation prevents the original query from diverging. In addition, restructuring of the logical terms and numerical constraints in queries served as an effective way of constraint relaxation for mutated queries to respond to the situation of no documents retrieved. Moreover, adding feedback to the system helps it to cater to the needs of the user more closely. It also helps to maintain a converging query trend.

It was also found that the efficiency of our genetic algorithm initially increases, reaches a maximum value, and then decreases as population size increases. This is the same for the number of generations. Thus, there is an optimal value of population size and number of generations when genetic algorithm is applied in query optimization. Also, by having longer queries, the real power of genetic algorithm will be unleashed.

Learning of Synonyms

With the usage of synonyms, it is rather tedious and difficult for a user to add synonyms to his ontology terms. Thus, there is a need to automate the learning of synonyms from other agents. The synonyms table of each agent may start off as one with only a few synonyms. Since the synonyms table is an ontology, learning of synonyms can be achieved by ontology exchange. During the interaction among agents, the agents will exchange their synonyms for a

particular query or even during product ontology exchange. For example, Agent1 requests for “Television” and in its synonyms table, “Television” only has 1 synonym “TV.” Agent2 received queries from Agent1 and in its synonyms table, “Television” can be found as either the ontology term or as one of the synonyms. It has the term “TV Set” as a synonym. Now, Agent2 recognizes that Agent1 does not have “TV Set” in its table. Thus, the term “TV Set” is added to the synonyms table of Agent1. In this way, the synonyms table can be expanded easily as more interaction is made.

Training the Genetic Algorithm

As mentioned, there is an optimal value of population size and number of generations when genetic algorithm is applied in query optimization. These optimum values would probably depend on the type and size of database. However, in most situations, it is quite impossible to predict the type of database that the system will be accessing. So it is suggested that the genetic algorithm undergo some training such that fine tuning of these parameters can be achieved when they revisit the same database.

Nevertheless, some users might wish to trade off some efficiency for that extra bit of optimization. Thus, the system should also allow some tolerance for the user to decide.

Improvement in Feedback

The feedback allows more interaction with the user. There are still many ways in which the environment can be made more friendly. For example, during the feedback to the user, if the user finds that too many documents would be retrieved, it will be better if the system can allow the user to add in more requirements to his original query. At the same time, the user can delete some requirements from his original query if too few documents are retrieved.

REFERENCES

- Boughanem, M., Chrisment, C., & Tamine, L. (1999). Genetic approach to query space exploration. *Information Retrieval*, 1(3), 175-192.
- Braga, R. M. M., Werner, C. M. L., & Mattosso, M. (2000, September 6-8). Using ontologies for domain information retrieval. *Proceedings of the 11th International IEEE Conference on Database and Expert Systems Applications*, Greenwich, London, UK (pp. 836-840).
- Fensel, D. (2000, November). The Semantic Web and its language. Trends & controversies. *IEEE Intelligent Systems*, 15(6), 67-73.
- Hendler, J. (2001). Agents and the Semantic Web. *IEEE Intelligent Systems*, 16(2), 30-37.
- Karp, P. D. (2000). An ontology for biological function based on molecular interactions. *Bioinformatics*, 16(3), 269-285.
- Kouichi, A. B. E., Toshihiro, T., & Hiroshi, N. (1999, September 21-24). An efficient information retrieval method in WWW using genetic algorithm. *Proceedings of the International Workshop on Parallel Processing*, Wakamatsu, Japan (pp. 522-527).
- Kraft, D. H., Petry, F. E., Buckles, B. P., & Sadasivan, T. (1994, June 27-July 2). The use of genetic programming to build queries for information retrieval. *Proceedings of the 1st IEEE Conference on Computational Intelligence*, Florida (pp. 468-473).
- Losee, R. M. (1991). An analytic measure predicting information retrieval system performance. *Information Processing and Management*, 27(1), 1-13.
- Losee, R. M. (1998). Parameter estimation for probabilistic document retrieval models. *Journal of the American Society for Information Science*, 39(1), 1-16.
- Losee, R. M. (2000, July 2). Learning syntactic rules and tags with genetic algorithms for information retrieval and filtering: An empirical basis for grammatical rules. *Information Processing & Management*, 32(2), 185-197.
- McGuinness, D. L. (2001). Ontological issues for knowledge-enhanced search. *Proceedings of FOIS'98*, Italy (pp. 302-316). Washington, DC: IOS Press.
- Robertson, S. E., & Sparck-Jones, K. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27(3), 129-146.
- Rzhetsky, A., Tomohiro, K., Sergey, K., Shawn, M. G., Michael, K., Sabrina, H. K et al. (n.d.). *A knowledge model for analysis and simulation of regulatory networks*. Retrieved from http://www.columbia.edu/cu/cie/GeneWays_fact_sheet.pdf
- Salton, G. (1989). *The transformation, analysis and retrieval of information by computer*. Reading, MA: Addison-Wesley.
- Salton, G., & McGill, M. (1983). *Introduction to modern information retrieval*. New York: McGraw-Hill.
- Wieerhold, G., Stefan, D., Sergey, M., Prasenjit, M., Yuhui, J., Sichun, X., et al. (2000). *OntoAgents—A project in the DARPA DAML PROGRAM*. Retrieved from <http://www-db.stanford.edu/OntoAgents>
- Yang, J. J., & Korfhage, R. R. (1994). Query modification using genetic algorithms in vector space models. *International Journal of Expert Systems in Research Applications*, 7(2), 165-191.

Chapter XXXVII

Intelligent Product Brokering Services

Sheng-Wei Guan
Brunel University, UK

ABSTRACT

Agent-based system has great potential in the area of m-commerce and a lot of research has been done on making the system intelligent enough to personalize its service for users. In most systems, user-supplied keywords are normally used to generate a profile for each user. In this chapter, a design for an evolutionary ontology-based product-brokering agent for m-commerce applications has been proposed. It uses an evaluation function to represent the user's preference instead of the usual keyword-based profile. By using genetic algorithms, the agent tries to track the user's preferences for a particular product by tuning some of the parameters inside this function. A Java-based prototype has been implemented and the results obtained from our experiments look promising.

INTRODUCTION

In this age of information technology, there has been an increasing demand for more and more sophisticated software that are capable of integrating and processing information from diverse sources. Traditional software technologies have failed to keep pace with these increasing demands, and alternative solutions are being considered. Agent-based systems (Nwana

& Ndumu, 1996; Aylett, Brazier, Jennings, Luck, Preist, & Nwana, 1998) have been proposed as a potential solution, and much research has been done on this relatively new technology.

One of the potential applications for agent technology is in the area of m-commerce. According to a study done by Frost and Sullivan¹, it has been projected that electronic commerce conducted via mobile devices such as cellular

phones and PDAs (personal digital assistants) will become a whopping \$25 billion market worldwide by 2006. Some of the driving factors behind the m-commerce “revolution” have been attributed to the compactness and high penetration rate of these mobile devices. This, along with the relatively low cost of entry for most service providers, has made m-commerce the buzzword of the next century.

CURRENT SITUATION AND MOTIVATION OF RESEARCH

However, despite all the hype and promises about m-commerce, several main issues (Nwana & Ndumu, 1997; Morris & Dickinson, 2001) will have to be resolved before agent technology can be fully adopted into any m-commerce systems. Clumsy user interfaces, cumbersome application, low speeds, flaky connections, and expensive services have soured many who have tried m-commerce. In fact, a usability study done in London by the Nielsen Norman Group² has found that about 70% of the participants have said that they would not want to use a WAP- (wireless application protocol) enabled phone again within a year, after they tried it for a week. Security and privacy concerns have also dampened enthusiasm for m-commerce. One of the concerns has been the fact that mobile devices such as PDAs are very easy to lose. They are also an easy prey for thieves, and unauthorized personnel can have easy access to the valid user ID and passwords stored in these devices to make fraudulent transactions.

Taking all these concerns into account, it seems like good old e-commerce will remain as the preferred choice for online transactions for many years to come. Customers will only use wireless mobile device to access the Internet if they have a good reason to do so. Therefore, in

order to entice customers to participate in m-commerce, the developers will have to offer something that is unique and which no self-respecting consumer can live without. One of the potential “killer” applications for m-commerce could be an intelligent program that is able to search and retrieve a personalized set of products from the Internet for its user.

Currently, when a user wants to search for a particular product on the Internet, what he will normally do is to use popular search engines such as Altavista³ or Yahoo!⁴, and enter keywords that describe the product. These search engines will process these keywords and churn out a large number of links for the user to visit. On the other hand, if the user already knows of some URLs that might have the product information, he will visit these Web sites and hopefully get the information that he is looking for.

Although these are the more common methods of searching for information on the Internet, it need not necessarily be the best or the most efficient ones. Neither the search engine nor the Web site knows the preference of the user and hence might provide information that is totally irrelevant to the user. For example, if the user wants to search for information about “software agents,” the search engine could return links to “insurance agents” instead. A significant amount of time could be wasted on such irrelevant information which could have been better spent on other, more important tasks.

In an agent-based m-commerce, agents act on behalf of their users by carrying out delegated tasks automatically. Currently, there is no single agent that can perform all the tasks meted out by the user. Like humans, specialized agents are required that are able to work in a specific type of environment. A product brokering agent seems to be a potential solution for this scenario. The agent will search for the

products in the background with minimal user intervention, thereby allowing the user to concentrate on other tasks. It could be programmed with the user's preferences in mind and filter out irrelevant products automatically. The agent could also detect shifts in the user's interest and, through some evolution mechanism, adjust itself accordingly to suit the user.

OVERVIEW OF THE CHAPTER

This chapter describes the design of an intelligent ontology-based product brokering agent capable of providing a personalized service for its user. It does this through *user profiling* (Soltysiak & Crabtree, 1998). Such agents are able to learn the preferences of the user over time and recommend products that might interest the user. The agent achieves this either by user feedback or through its own observation. This technique has been used quite successfully for specific types of agent tasks, typically those that are information intensive and often involve the World Wide Web (WWW).

We first highlight some of the related works that have been done by other researchers. A proposed design for an evolutionary product brokering agent will then be presented to the reader. A prototype of the product brokering agent has been implemented using Java, and the system is put through a series of tests. Results obtained from these tests are discussed later in this chapter. Although the results are encouraging, some limitations of the system will also be highlighted. Potential applications for a product brokering agent in m-commerce will then be discussed. Finally, the last section presents some concluding remarks along with a discussion on the possible extensions to the current work.

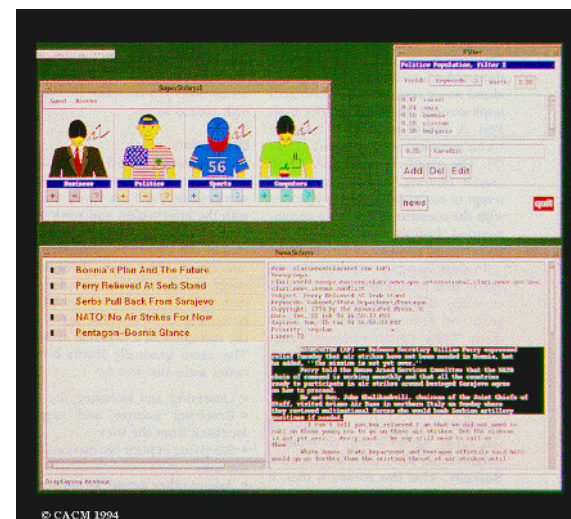
BACKGROUND

Personalized product brokering agents require a profile of the user in order to function effectively. The agent would also have to be responsive to changes in the user's interests, and be able to search and extract relevant information from outside sources. The rest of this section will highlight some of the works done by other researchers which are closely related to a product brokering agent.

At MIT Media Labs, Sheth and Maes (1993) have come up with a system that is able to filter and retrieve a personalized set of USENET articles for a particular user. This is done by creating and evolving a population of information filtering agents using genetic algorithms (Zhu & Guan, 2001). A screenshot of their implemented system is as shown in Figure 1.

Some keywords will be provided by the user representing the user's interests. Weights are also assigned to each keyword, and the agents will use them to search and retrieve articles from the relevant newsgroups. After reading the articles, the user can either give positive or

Figure 1. Screenshot of the information filtering system



negative feedback to the agents via a simple GUI. Positive feedback increases the fitness of the appropriate agent(s) and also the weights of the relevant keywords (vice versa for negative feedback). In the background, the system periodically creates new generations of agents from the fitter species while eliminating the weaker ones. Initial results obtained from their experiments have been encouraging and showed that the agents are capable of tracking its user's interests and recommend mostly relevant articles.

While the researchers at MIT require the user to input their preferences into the system before a profile can be created, Soltysiak and Crabtree (1998) believed that the user's profile can be generated automatically by monitoring the user's Web and e-mail habits, thereby reducing the need for user-supplied keywords.

Their approach is to extract high information-bearing words, which occurs frequently in the documents that are opened by the user. This is achieved by using ProSum⁵, which is a text summarizer that can generate a set of keywords to describe the document and also determines the information value of each keyword. A clustering algorithm is then employed to help identify the user's interests, and some heuristics are used to ensure that the program could perform as much of the classification of interest clusters as possible, thereby minimizing the amount of user input required in the profile generation process.

However, they have not been completely successful in their own experiments. The researchers admitted that it would be very difficult for the system to classify all of the user's interest without the user's help. Nevertheless, they believed that their program has taken a step in the right direction by learning user's interest with minimal human intervention.

A new product brokering agent usually does not have sufficient information to recommend any products to the user. Hence, it has to get

product information from somewhere else. A good source of information will be the Internet. In order to do that, a method suggested by Pant and Menczer (1998) is to implement a population of Web crawlers called *InfoSpiders* that searches the WWW on behalf of the user. It will gather information on the Internet based on the user's query and indexes them accordingly. It behaves much like a personalized search engine, but is designed to evolve and retrieve only relevant Web pages for its user.

These agents initially rely on traditional search engines to obtain a starting set of URLs, which are relevant to the user's query. The agents will then visit these Web sites and decode their contents before deciding where to go next. The decoding process includes parsing the Web page, and by looking at a small set of words around each hyperlink, a score is given based on their relevance to the user. The link with the highest score is then selected and the agent visits the Web site.

No further details have been provided on how they extract or analyze the contents of the Web pages, but it has been mentioned that they use neural networks, HTML, and XML parsing tools that are commonly used by other Web crawlers. The agent stops after they had visited a pre-determined number of Web pages or when it could no longer find any relevant Web pages. The user can also terminate the search anytime he wishes while the program is still running.

DESIGN OF PRODUCT BROKERING AGENT

A product brokering agent can be used to search for all kinds of products. In our application, the agent will be used to search for some computer products, namely *CPU*, *Mainboard*, and *Memory*. It is possible to extend the application to search for other products. All the

codes are written in Java, as it is object oriented in nature and is compatible across multiple operating systems.

Similar to the information filtering agents done by Sheth and Maes (1993), an initial population of product brokering agents will be created and evolved using some form of genetic algorithms. However, in this design, the *profile* of the user is not based on any keywords supplied by the user. In fact, no keywords are required to be entered by the user. Instead, each agent will have an evaluation function that will be used to calculate the *value* of each product. Products that have a higher *value* will have a higher chance of being recommended by the agent. This evaluation function has some tunable parameters, which characterizes the user's preferences for a particular category of products. Initially, these tunable parameters will be randomly generated based on some heuristics, but they will evolve over time to match the user's preferences.

In this design, some assumptions have been made about the system. One of the most important assumptions is that the user of the system is a rational person and will select a product rationally. Another important assumption is that the value, which a user places on a product, can be calculated mathematically. The product values that we are focusing on will be those that can be calculated by using some tangible attributes (e.g., *price*) of the product. The agent will not be able to calculate the intangible value (e.g., *branding*) that a user has placed on the product. If these assumptions are not met, the agent will not be able to track the user's preferences successfully.

Ontology

Before the product brokering agent is able to explore the Internet and retrieve product information for the user, the agent needs to have some prior knowledge such as the URL of

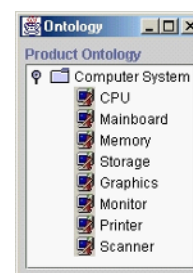
some relevant Web sites, keywords, or quantifiable attributes that can be used to describe the product. It could be very tedious if the user has to enter such information into the agents when he wants to search for a particular product. Imagine the amount of data he will have to enter if he wants to search for several different products.

The *product ontology* has been implemented in a tree-like structure, with the leaf nodes representing the products and the parent node representing the product category. Each leaf node actually contains a Java class called *productInfo*, which has some prior information about the product. Different products will have a different *productInfo* class. New products can be added as a leaf node to the parent node easily. When the leaf node is selected, it will pass the product information to the product brokering agents. Currently, selection of the leaf node will pass the URL of the product's Web site and its attributes to the agents automatically.

Product Brokering Agent

After describing how the agents are going to obtain their product knowledge, the next stage is to define the agent itself. The agent will basically be a Java programming thread that will be running continuously in a while-loop until some terminating conditions have been met. A unique agent name will be given to each agent

Figure 2. Screenshot of product ontology



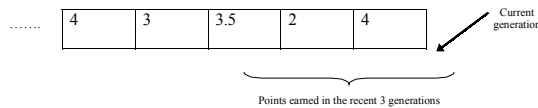
so that we can identify and differentiate the agents from one another.

Agent's Fitness Function

To calculate the fitness of the agent, the proposed fitness function has been defined by using the following equation:

$$\text{Fitness} = \frac{\sum_{n=\text{window\_size}}^{\text{points earned in the recent } n \text{ generations}}}{n} \quad (1)$$

This fitness function basically sums up and calculates the average amount of points earned by the agent in the current and the previous $n-1$ generations, where n is the window size of the agent's short-term memory. For example, assume an agent has a *window_size* of 3 and the following diagram shows the amount of points earned in each generation:



In this example, the total amount of points earned by the agent in the recent three generations is 9.5. Therefore the fitness of the agent is about $\frac{9.5}{3} = 3.1667$. The amount of points earned before that will not be considered. The rationale for this is that more emphasis should be placed on the agent's current performance instead of its past performances. As the fitness of an agent would be used to determine which agent to evolve, we do not want its past performances, which might be irrelevant now, to influence the evolution process.

An agent's fitness will always be a positive value, and a new agent would start off with

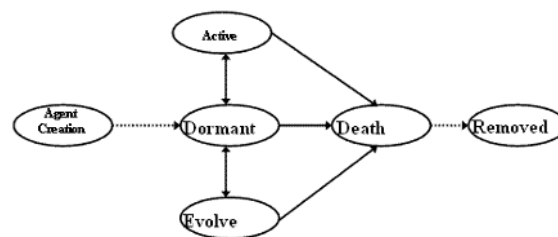
some default fitness. The fitness of an agent can also be a good indicator about the agent's performance. Therefore, in order to keep track of the agent's performances, each agent will have a list called *fitness_history*, which is used to store the fitness of an agent for each generation. Hence, after an agent has been awarded some points, it will calculate its new fitness using Equation 1 and insert the value into the *fitness_history* list. Details on how an agent earns its points will be discussed in the next few sections.

Agent's Lifecycle

Once an agent has been created, it can be in any of the four different states, namely *Dormant*, *Active*, *Evolve*, and *Death*. During agent creation, the agent will register itself to a database and its default state will be the *Dormant* state. It would also de-register itself before it is removed from the system. This is to allow the user to keep track of all the agents running in the system. The database also allows the agents to store any product information that they have found on the Internet. Figure 3 shows how the four different states are implemented.

Dormant—The agent is not doing any task at the moment. It is waiting for the user to give it instructions. Note that this will be the default state of the agent once it has been created. At this state, the user can modify the agent's parameters before starting the agent.

Figure 3. Agent's lifecycle



Active—The agent has received some instructions and is currently performing some tasks for the user. The types of tasks performed by the agents will be discussed further on.

Evolve—The agent has received some user feedback regarding its performance. It is now analyzing this feedback and making the appropriate adjustments.

Death—The agent has been killed and it can no longer perform any tasks for its user. Note that it is still present in the system and will only be removed completely when instructed by the user. This allows the user to recycle any information that he might find suitable.

Agent's Task

Once the user has passed some instructions to the agent, it will switch to the *Active* state and activate the appropriate tasks. As an agent might need to perform different types of tasks simultaneously, these tasks are implemented as independent and self-contained programs, which are separated from the agent. Therefore, instead of implementing several agents from scratch to perform different tasks, we only need to implement a basic agent and a few task programs. What the basic agent needs to do is to call the appropriate task program, pass some information to it, and the task programs will handle the rest.

For our application, the agent's task has been designed specially to parse information from Hardwarezone.com⁶, a Web site hosted in Singapore that displays up-to-date information of various computer products in table form. The task program allows the agent to establish a connection to the Web site and download the HTML document onto a local computer. The program then parses the document and extracts the relevant information for the agent by looking for specific tags within the HTML docu-

Figure 4. Screenshot of an agent's database

Description	Performance	Price	Shop	End value
Intel Pentium	1.65	322 C-sec	2.000000	2.000000
Intel Pentium	1.3	200 C-sec	2.700000	2.700000
Intel Pentium	1.65	305 C-sec	2.710000	2.710000
AMD Athlon XP	1.65	211 C-sec	2.800000	2.800000
Intel Pentium	1.3	388 C-sec	2.850000	2.850000
Intel Pentium	1.7	247 C-sec	2.850000	2.850000
Intel Pentium	1.5	199 C-sec	2.850000	2.850000
Intel Pentium	1.3	410 C-sec	2.850000	2.850000
AMD Athlon XP	1.4	110 C-sec	2.850000	2.850000
AMD Athlon XP	1.45	110 C-sec	2.850000	2.850000
Intel Pentium	1.3	450 C-sec	2.850000	2.850000
AMD Athlon XP	1.7	312 C-sec	2.850000	2.850000
Intel Pentium	1.7	312 C-sec	2.850000	2.850000
AMD Athlon 1.4	1	115 C-sec	2.850000	2.850000
Intel Pentium	1.5	270 C-sec	2.850000	2.850000
Intel Pentium	1.3	450 C-sec	2.850000	2.850000
Intel Pentium	2.65	100 C-sec	2.850000	2.850000
AMD Athlon XP	1.65	305 C-sec	2.850000	2.850000
Intel Pentium	1.3	450 C-sec	2.850000	2.850000
AMD Athlon XP	1.65	250 C-sec	2.850000	2.850000
AMD Athlon 1.3	1.25	111 C-sec	2.850000	2.850000
Intel Pentium 1.1	1	670 C-sec	2.850000	2.850000
AMD Athlon 1.2	1.47	278 C-sec	2.850000	2.850000
AMD Athlon 1.2	1.25	89 C-sec	2.850000	2.850000

ment. In our application, the program will be able to extract information such as the description of the product, its price, and the name of the shop that is selling this product.

Agent's Knowledge

After an agent has retrieved some product information, it needs a place to store this piece of information. As mentioned, when an agent is created, it will register itself to a database. A Microsoft Access database is used in this application. Within the database, a table will be created for each agent to store all the information that it has retrieved. In addition, the agents will also store this data on a *global* database. The *global* database will contain all the products that have been retrieved by the agents in the system. Figure 4 shows a screenshot of an agent's database.

Product Recommendation

Before recommending a product to the user, the agent should first be able to evaluate which product would best fit the user's requirements. A proposed method is to use some quantifiable attributes such as *cost*, *performance*, and so forth to evaluate the products. An example of

an evaluating function could be the following equation:

$$product\_value = perf\_weight * performance - cost\_weight * cost \quad (2)$$

The attributes used in Equation 2 try to model the two types of factors that can influence a user's choice. The first attribute (*performance*) represents the performance of the product, while the second attribute (*cost*) represents the cost of the product. It has been assumed that the better the product, the higher will be its *performance*, and a better product usually results in a higher *cost*. From Equation 2, it can be seen that a product with a higher *performance* and/or a lower *cost* will result in a higher *product_value*.

The two weights *perf_weight* and *cost_weight* represent the weights that the user could give to each attribute. These two parameters are actually used to represent the user's preferences and are incorporated inside the agent. If *perf_weight* has a higher value, it means that the user places more emphasis on the performance of the product. Likewise, if the user has a higher value for *cost_weight*, it means that the user is more concerned about the cost of the product. Note that, for different products, a different set of attributes and weights could be defined for Equation 2, and all these could be defined inside the *product ontology*.

When an agent is created, these two weights will be initialized based on some heuristics and would be used to calculate the value of each and every product found in the agent's database. The agent will then rank the products according to their values and select the top three products to be presented to the user. The value of *perf_weight* and *cost_weight* will be allowed to change when the agent undergoes evolution.

Agent's GUI

It will be useful if the user is able to observe what is happening inside an agent when required. To facilitate this, each agent will have a simple GUI that shows information such as the name of the agent, its current status, products recommended, and so forth. It would also allow the user to change some of the parameters inside the agents.

The agent's GUI is implemented as shown in Figure 5. The GUI allows the user to see what is the top product inside the agent's database and also some of its internal parameters. The user can kill the agent from this GUI by clicking the *kill* button or update some of the agent's parameters by using the *update* button. When the user clicks the *history* button, the agent's history will be shown to the user (Figure 6). The *database* button will pop up a GUI (similar to Figure 4) which shows all the products inside an agent's database.

Figure 5. Agent's GUI

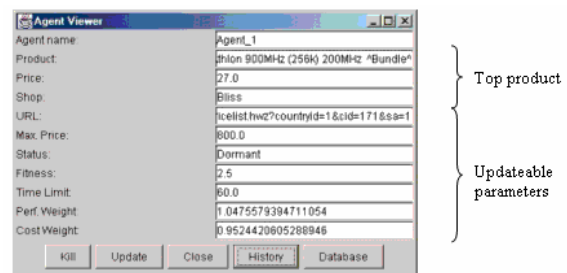
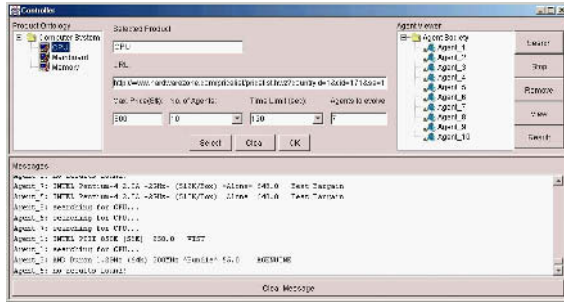


Figure 6. Agent's history

Generation	Fitness	Perf_weight	Cost_Weight
Generation 0 ->	2.5	1.047557	0.952442
Generation 1 ->	1.25	1.047557	0.952442
Generation 2 ->	1.70	1.047557	0.952442
Generation 3 ->	2.16	1.047557	0.952442

Figure 7. Screenshot of the monitoring tool



Monitoring Tools

A monitoring tool will be provided for the user which allows him to observe and control the behavior of his agents while they are searching for products on the Internet. This tool will be the main interface between the user and his agents.

The user can choose from a list of products provided in the *product ontology* and enter some parameters (e.g., number of agents, etc.) before starting the search. Once all the parameters have been entered into the system, the appropriate number of agents will be created to search for the product on the Internet. While the program is running, the user can start, stop or provide feedback to the agents anytime by using this tool. A text message will also be provided to allow the user to track the progress of the agents in the system. A screenshot of the implemented system is as shown in Figure 7.

User Feedback

During user feedback, each agent in the system will select the top three products in its database and add them into a *recommended* list. A sorting function will be implemented to allow the user to sort the list according to his preferences. If the user cannot find any product that he fancies in this list, he can look at the *global* list, which contains all the products that have

Figure 8. Screenshot of an agent's database after sorting

No.	Description	Price	Shop
1	AMD Athlon XP 1900+ (1.6GHz, 256MB) = Bundle	288.0	BiSS
2	INTEL Pentium 4 1.70GHz/256KB = Bundle	289.0	Casstronic
3	INTEL Pentium 4 1.70GHz/256KB = Alone	325.0	Byssnet
4	AMD Athlon MP 1500+ (1.33GHz, 256MB) = Alone	330.0	MediaPro
5	INTEL PM 1.30GHz (Tulatro) - Box "Alone"	340.0	MediaPro
6	INTEL PM 1.26GHz (Tulatro) - Box "Alone"	399.0	MediaPro
7	INTEL Pentium 4 1.80GHz = Bundle	399.0	Video-Pr
8	INTEL Pentium 4 1.80GHz = Alone	418.0	Cybernet
9	INTEL Pentium 4 1.80GHz = Alone	439.0	MediaPro
10	INTEL PM 1.26GHz (Tulatro) - Box "Bundle"	445.0	Taknet
11	INTEL Pentium 4 1.80GHz = Bundle	452.0	BiSS
12	AMD Athlon MP 1900+ (1.6GHz, 256MB) = Bundle	480.0	Superpet
13	AMD Athlon MP 1800+ (1.53GHz, 256MB) = Alone	495.0	BiSS

been retrieved by the agents in the system. When the user sees a product that he likes, he can select the product by clicking on it and all the agents in the system will be informed about the user's selection.

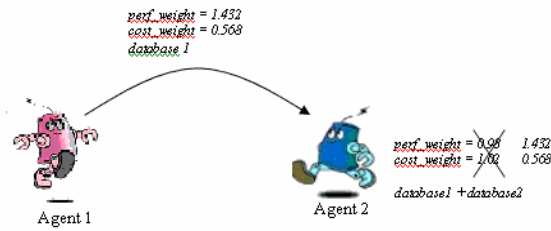
The agents will take note of the product that the user has selected and search for that product inside its own database. At this stage, each agent would have already assigned a product value to each and every product in its database. To determine the amount of points to award to an agent, it will be asked to rank the products in an ascending order according to this value. Hence, products with a higher value will be located at the bottom of the table. The agent will then determine the position of the user-selected product and take note of its row number. The formula to calculate the exact amount of points to give to an agent is as follows:

$$\text{points awarded} = \frac{\text{row number of user selected product}}{\text{total number of products}} \times \text{maximum points} \quad (3)$$

As an example, assume Figure 8 shows the agent's product list after it has been sorted in ascending order.

In this example, there are a total of 13 products found inside the agent's database, and the product with the highest product value is located at row 13. This will be the top product inside the agent's database. However during

Figure 9. Evolution process



feedback, the user might have actually chosen the product at row number 7 instead. Therefore, assuming that 5 is the maximum amount of points awardable, the amount of points that the agent earns in this case will be:

$$points\ awarded = \frac{7}{13} \times 5 = 2.692$$

Using this example, it can be seen that if the rankings of the two selected products are far apart, the agent will actually receive less points. Also, if the user-selected product is not inside the agent's database, the agent will not receive any points at all!

Evolution Process

The fitness of an agent will be used to decide which agent will undergo the evolution process. In conventional genetic algorithm, the agent with a higher fitness will have a higher chance of survival as compared to an agent with a lower fitness. However, in this application, there will be a slight variation in the algorithm. Instead of killing the weaker agents, they will simply copy over all the parameters of the fitter agents. The weaker agent will also copy over the database of the fitter agent and merge the two databases together to form a larger database. Let *Agent 1* be the fitter agent and *Agent 2* be the weaker one. Figure 9 shows what happens between the agents during this evolution process.

However, the parameters inherited by the new agent in this evolution process might not necessarily be the most optimal. Therefore, the new agent will try to adjust its newly acquired *perf_weight* and *cost_weight* to better reflect the user's requirements. First, it will use the newly acquired parameters to re-evaluate all the products found inside its new database. Then the agent will select the best product based on these new parameters. If it is the same as the user-selected product, no further changes will be required, but some small and random mutations in the parameters will be allowed.

However, that will not usually be the case. In this case, the agent will compare the *performance* and *cost* attributes of the products that are selected by the user and the agent. Let $p1$ and $p2$ denote the *performance* of the products selected by the user and agent respectively. Also let $c1$ and $c2$ denote the *cost* of the products selected by the user and agent. Four possible scenarios will have to be considered:

1. $p1 > p2$ and $c1 > c2$
The user has selected a product that has a much better performance but more expensive than what the agent has suggested. The agent can deduce that the user places more emphasis on the performance rather than the cost of the product. Therefore, it will increase its *perf_weight* and reduce its *cost_weight*.
2. $p1 < p2$ and $c1 < c2$
The user has selected a product that is of a lower performance but cheaper than what the agent has suggested. The agent can deduce that the user places more emphasis on the cost rather than the performance of the product. Therefore, it will reduce its *perf_weight* and increase its *cost_weight*.

Figure 10. Recommended list of products

No.	Description	Price	Shop
1	AMD Athlon 1 1000+ (2...)	155.0	IMS Systems
2	AMD Athlon 1 2016 (2...)	59.0	Oliss
3	AMD Athlon XP 1800+ ...	385.0	IMS
4	AMD Athlon XP 1800+ ...	430.0	Superdeal
5	AMD Athlon XP 1800+ (L...)	122.0	no Data
6	AMD Athlon XP 1800+ (L...)	130.0	no Data
7	AMD Athlon XP 1800+ (L...)	295.0	no Data
8	AMD Athlon XP 2300+ (L...)	475.0	Shynet
9	AMD Athlon XP 2300+ (L...)	309.0	Video-Francis
10	AMD Duron 800MHz (6...)	69.0	Laser
11	INTEL Celeron 400MHz ...	100.0	RFM
12	INTEL Celeron 400MHz ...	85.0	Video-Francis
13	INTEL Pentium 4 1.8G ...	379.0	MediaPro
14	INTEL PIII 750E (550...)	135.0	HardwarePlace
15	INTEL PIII 750E (550...)	155.0	IMS

3. $p1 < p2$ and $c1 > c2$

The user has selected a product that is of a lower performance and more expensive than what the agent has suggested. The agent will be confused over such a selection and will prompt the user if it should carry on the evaluation. If the user still wants the agent to carry on, it will either reduce its *perf_weight* or *cost_weight*. This might happen when the user has placed some form of intangible attributes/values on the product which are not present inside the agent's evaluation function.

4. $p1 > p2$ and $c1 < c2$

The user has selected a product that is of a higher performance and cheaper than what the agent has suggested. This scenario will not arise during evolution. Looking back at Equation 2, a product with a higher performance and/or a cheaper product will result in a higher value being assigned to that product. Since using $p1$ and $c1$ will definitely result in a higher product value as compared to $p2$ and $c2$, this scenario will not happen.

The evolution process as described in this section is actually quite similar to the use of the reproduction and crossover operator to clone the fitter agents and then use the mutation operator to mutate some of the parameters within the agents.

SYSTEM EVALUATION

To evaluate the performance of the implemented system, some simple experiments have been conducted to see if the agents are able to track the user's preference. All the agents in these experiments are running inside a single computer. The computer used in this experiment is a Pentium 4 with 384MB of memory and operating under the Windows ME environment with JDK1.3.0. The system connects to the Internet via a 56.6kbps modem.

Product Recommendation

In this experiment, a group of 20 randomly generated product brokering agents are instructed to search for one of the products on the Internet. The product chosen for this experiment is the *CPU*. For this part of the experiment, the user wants to get the best *CPU* possible and he does not mind the price. After instructing the agents to search for the product, the system is allowed to run on its own for about 10 minutes so that the agents can retrieve sufficient products before the user gives feedback. After 10 minutes, the user clicks on the *result* button and the recommended list is as shown in Figure 10.

From the recommended list, the user selects the current best product at row 13, which happens to be Pentium 4 1.8GHz as shown in Figure 11.

Figure 11. User selection

No.	Description	Price	Shop
1	AMD Athlon 1 1000+ (2...)	155.0	IMS Systems
2	AMD Athlon 1 2016 (2...)	59.0	Oliss
3	AMD Athlon XP 1800+ ...	385.0	IMS
4	AMD Athlon XP 1800+ ...	430.0	Superdeal
5	AMD Athlon XP 1800+ (L...)	122.0	no Data
6	AMD Athlon XP 1800+ (L...)	130.0	no Data
7	AMD Athlon XP 1800+ (L...)	295.0	no Data
8	AMD Athlon XP 2300+ (L...)	475.0	Shynet
9	AMD Athlon XP 2300+ (L...)	309.0	Video-Francis
10	AMD Duron 800MHz (6...)	69.0	Laser
11	INTEL Celeron 400MHz ...	100.0	RFM
12	INTEL Celeron 400MHz ...	85.0	Video-Francis
13	INTEL Pentium 4 1.8G ...	379.0	MediaPro
14	INTEL PIII 750E (550...)	135.0	HardwarePlace
15	INTEL PIII 750E (550...)	155.0	IMS

Figure 12. Recommended list after feedbacks

No.	Description	Price	Shop
1	AMD Athlon XP 2000+ (	475.0	Sysnet
2	AMD Athlon XP 2000+ (	309.0	Video-Pro.com
3	INTEL Pentium 4 1.8A...	379.0	MediaPro
4	INTEL Pentium 4 1.9C...	452.0	Diss
5	INTEL XEON 1.70GHz (	520.0	IMS

While the feedback is been made, the system continues to search for products in the background. After making a few similar selections, the agents evolved and re-evaluated their list. The new recommended list is now as shown in Figure 12. The list now only shows the best *CPU* that has been retrieved by all the agents.

When the user is satisfied with what the system has learned, he allows the system to carry on searching the Internet for new products on its own. After some time has passed, the agents have found an even better performing *CPU* and it is reflected in the agent's recommended list.

Two other scenarios have also been tested on the system. One of them is to search for the cheapest *CPU* available, while the other is to find a right mix of *performance* and *cost* for the user. The steps used in these scenarios are similar to those used in the first part of the experiment, and the results obtained are encouraging.

Tracking User's Preferences

In this experiment, the objective is to test if the system is able to detect a change in the user's preferences. If it is able to do so, we need to determine how fast the system will be able to respond to these changes. This could be observed by looking at the average fitness of all the agents in the system. The average fitness of all the agents should remain high if the system

Figure 13. New products recommended by the agents

No.	Description	Price	Shop
1	AMD Athlon XP 2000+ (1	475.0	Sysnet
2	AMD Athlon XP 2000+ (1	309.0	Video-Pro.com
3	INTEL Pentium 4 1.8A-1	379.0	MediaPro
4	INTEL Pentium 4 1.9C...	375.0	Video-Pro.com
5	INTEL Pentium 4 1.8A...	463.0	Bitco
6	INTEL Pentium 4 2.0A-2	545.0	Global IT Mart
7	INTEL XEON 1.70GHz (25	520.0	IMS

is able to track and respond to the changes effectively.

An initial population of 20 randomly generated agents is created, and the response of the system is observed by changing the number of agents to evolve in the population. For each test case, the same set of test data are used and will be described in the next few sections.

Gradual Changes in User's Preferences

In this part of the experiment, the user starts by selecting the best *CPU* available. After a few selections, the user will gradually choose cheaper and cheaper *CPUs*. The experiment stops after all agents begin to recommend the cheapest *CPU* available. Figure 14 shows the

Figure 14. Tracking gradual changes in user's preferences

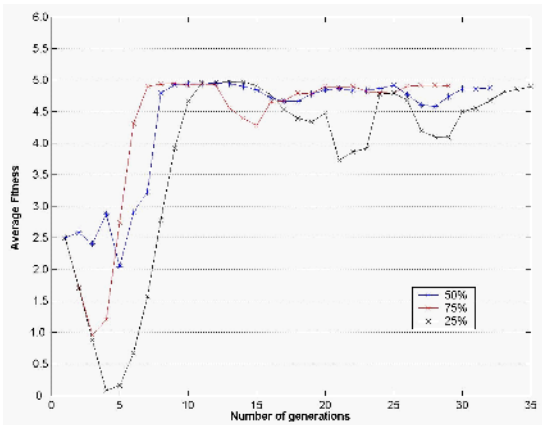
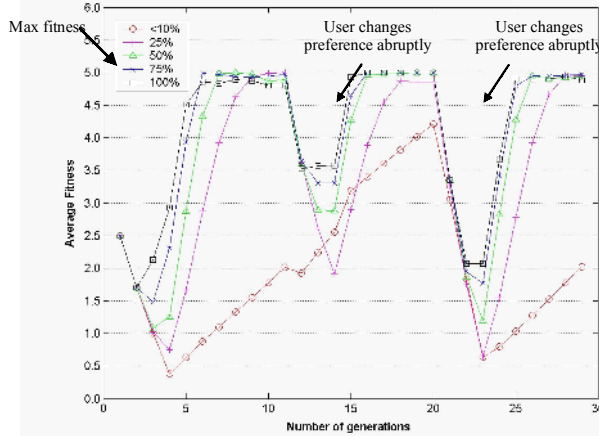


Figure 15. Tracking abrupt changes in user's preferences



average fitness of the agents, when the user gradually changes his preferences.

The results obtained from this experiment have shown that the system is capable of tracking gradual changes in the user's preferences. Although some "dips" are observed during the experiment, the average fitness of the agents in the system remains high while the user changes his selection. These dips could happen because some of the agents might not have the products that the user has selected in their database. Therefore, these agents do not receive any points and could "pull down" the average fitness quite significantly.

Abrupt Changes in User's Preferences

In this part of the experiment, the user makes two abrupt changes in his preferences. Initially,

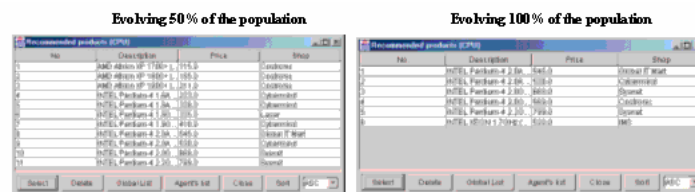
the user starts by selecting the best *CPU* available. After a while, he abruptly changes his preferences by selecting the cheapest *CPU* and then reverts back to his original selection. Figure 15 shows the average fitness of the agents when the user changes his preferences abruptly.

Interestingly, the results obtained suggested that as we increase the proportion of agents to evolve, the response of the system would be much better. The best result is obtained when 100% (black line) of the population is allowed to evolve. It can be seen that the system in this scenario attains the maximum average fitness in a shorter time as compared to others. Also, when the user makes an abrupt change in his selection, the average fitness of the system does not drop as much as compared to the rest, and it recovers much faster. The system could not track the user's preference when less than 10% (red line) of the population is allowed to evolve.

Although allowing a larger proportion of agents to evolve will result in a faster response, it might cause the system to converge to a solution prematurely. This has been observed when 100% of the population is allowed to evolve. The type of products recommended by the agents is not as diverse as compared to when only 50% of the population is allowed to evolve.

Hence, there is a tradeoff between them. From the experiment, it has been observed that allowing 50% of the population to evolve will be

Figure 16. Diversity of products recommended



a reasonable compromise between the response time and the diversity of products recommended.

Limitations of the System

Although preliminary test results have shown that the product brokering agents are quite successful in tracking the user's preferences after a few evolutions, it has been noticed that sometimes it does not follow the user's selection. A possible explanation for this could be due to the mismatch of the evaluation function that is used by the agent to calculate the perceived value of a particular product. Currently, the function only takes into account the performance and cost of a given product. However in reality, there might be other product attributes, which could also affect the user's decision. Some of them could also be intangibles, which might be difficult to represent in the function.

M-Commerce Applications

The proposed design of a product brokering agent has been implemented using Java in a desktop computer. However, mobile devices such as phones and PDAs tend to have a smaller screen, slower processors, as well as limited memory. Hence, this will pose some serious constraints when we want to transfer the software into these mobile devices. There is also a serious lack of standardization, as these mobile devices use different OS platforms, which makes it difficult for the developer to create a single program that can run on all devices.

After taking all these into consideration, a possible solution is to use software that is compatible across multiple operating platforms. A good candidate is Java, which has been used to implement the system as mentioned in this chapter. Java is a platform-independent software technology, allowing the same code to be

run on any system. This greatly reduces the time and cost of software development, and is particularly attractive for Internet or local network applications, because instructions can be sent over the Net without prior knowledge of the characteristics of the target device. However, the disadvantage of Java as compared to other programming languages, such as C, is that it is less efficient and has slower program execution. Faster processors and more memory are needed to compensate for this. This results in higher cost, and, for wireless applications, shorter battery life. However, this disadvantage has been slowly reduced by the introduction of more efficient JIT (just-in-time) compilers. Currently, the developers of Java have also introduced some highly optimized and micro versions of the Java software to cater specifically for small devices such as cellular phones and PDAs.

As for hardware, Philips Research⁷ has developed a co-processor that improves the execution of Java embedded software by a factor up to 10. This obviates the need for powerful processors and/or more memory that this programming language often requires, while maintaining the advantage of enabling fast and economic product development and easy integration with the Internet. The invention supports the use of Java and related languages in a rapidly growing and developing market, with applications ranging from smart cards to mobile phones and set-top boxes.

Application of Product Brokering Agent in M-Commerce

A PDA is an ideal device for m-commerce applications. It tends to have a larger screen and a more powerful processor as compared to a cellular phone, but is less bulky than a laptop. Making an existing application viewable in any wireless device—a process known as

Figure 17. Screenshot of a PDA with the implemented GUI



transcoding—is among one of the biggest challenges of m-commerce. In order to fit the screen of the PDA, the GUI implemented in this chapter will have to be scaled down to the appropriate size. A possible solution to fit all these into the PDA screen is to use scrollbars that allow the user to scroll the GUI. A possible screenshot of a PDA with the GUI is shown in Figure 17.

The PDA selected for our application is the Compaq iPAQ Pocket PC H3870. It has one of

the largest viewable screens in the market and also an integrated Bluetooth for wireless links to Bluetooth-enabled cellular phones. This device also supports the Java Virtual Machine, which will allow our software to be integrated into the PDA easily. The specifications of the PDA are shown in Table 1. For more detailed information, the reader is encouraged to visit the manufacturer's Web site⁸. A comparison of the iPAQ Pocket PC with other PDAs on the market can be found at PCWorld.com⁹.

CONCLUSION AND DISCUSSION

This chapter has demonstrated that by using genetic algorithm and an evaluation function, it is possible to design and implement an intelligent product brokering agent for m-commerce applications.

A simple prototype of the system has been implemented using Java, and the preliminary results obtained from the experiments have been encouraging. However, there are some limitations in the current prototype that might hamper the system's performance. More re-

Table 1. Specifications for Compaq's iPAQ Pocket PC H3870

iPAQ Pocket PC H3870

Operating System	Microsoft Pocket PC 2002
Processor	206 MHz Intel StrongARM 32-bit RISC processor
Display Type	Color reflective thin film transistor (TFT) LCD, 64K colors
Resolution/Viewable Image Size	240 x 320/ 2.26 x 3.02 inches
Pixel Pitch	0.24 mm
RAM	64MB
ROM	32MB
Input Method	Handwriting recognition, soft keyboard, voice record, inking
Wireless Connectivity	Bluetooth™, Infrared port (115 Kbps)
Dimensions	5.3" x 3.3" x .62"
Weight	6.7 oz. including battery

search will have to be done before a truly robust system can be implemented for m-commerce.

One of the possible improvements to the current work will be to allow the agents to be distributed in a network instead of being hosted entirely by the same computer. As the host computer might not have sufficient resources (processing power, bandwidth, etc.) to support all the agents in the system, it will be advantageous if some agents can be hosted by another computer. For m-commerce applications, this would mean the agents could now be hosted by a commercial Internet service provider (ISP).

An m-commerce user would not want to spend huge sums of money on maintaining a wireless connection to the ISP or a phone company. Likewise, it is unrealistic for mobile devices such as cellular phones and PDAs to be always "online." Currently, some ISPs do provide some form of storage spaces for their subscribers to store files inside their servers. In extension to this, an ISP could now also offer—with a reasonable fee—to host the agents that have been created by/for their subscribers. These agents could perform their tasks inside these servers and report back to its user when he re-establishes another connection with the ISP.

However, allowing agents to be distributed over the network will raise some issues, which the developer should look into before the system could be implemented. Since the agents are distributed, some form of communication protocol and *ACL* (Agent Communication Language) will have to be designed and incorporated into the system. One way is to upgrade the monitoring tool that has been implemented in this chapter so that it can communicate with the remote agents using some *ACL* via a socket connection. Currently, there exist some high-level agent languages such as *KQML* (Knowledge Query and Manipulation Language) (Finin, Fritzson, McKay, & McEntire, 1994) and *FIPA*

ACL, which has been developed for inter-agent communication.

Security issues also arise when agents are hosted by other computers. There is now a need for us to distinguish between agents that are sent by different users. In m-commerce, security is of paramount importance. Sensitive and private information of the user will have to be safeguarded from other hostile entities. This is especially important in the case of mobile agents. As they travel from host to host, we have to prevent them from being intercepted and its contents "core-dumped" by hostile hosts. Malicious agents could also masquerade as the original agent and trick an unsuspecting user into giving up his personal information.

REFERENCES

- Aylett, R., Brazier, F., Jennings, N., Luck, M., Preist, C., & Nwana, H. S. (1998). Agent systems and applications. *The Knowledge Engineering Review*, 13(3), 303-308.
- Finin, T., Fritzson, R., McKay, D., & McEntire, R. (1994). KQML as an agent communication language. *Proceedings of the 3rd International Conference on Information and Knowledge Management*.
- FIPA Specifications. (n.d.). Retrieved from <http://www.fipa.org/repository/fipa2000.html>
- Maes, P. (n.d.). *Agents that reduce work and information overload*. Retrieved from <http://pattie.www.media.mit.edu/people/pattie/CACM-94/CACM-94.p1.html>
- Morris, S., & Dickinson, P. (2001). *Perfect m-commerce*. London: Random House Business Books.
- Nwana, H. S., & Ndumu, D. T. (1996). An introduction to agent technology. *BT Technology Journal*, 14(4), 55-67.

Nwana, H. S., & Ndumu, D. T. (1997). Research and development challenges for agent-based systems. *IEEE Proceedings on Software Engineering*, 144(1), 2-10.

Pant, G., & Menczer, F. (2001). Evolve your own intelligent Web crawlers. *Autonomous Agents and Multi-Agent Systems*, 5(2), 221-229. University of Iowa.

Sheth, B., & Maes, P. (1993). *Evolving agents for personalized information filtering*. Boston: MIT Media Lab.

Soltysiak, S., & Crabtree, B. (1998). Automatic learning of user profiles—Towards the personalization of agent services. *BT Technology Journal*, 16(3), 110-117.

Soltysiak, S., & Crabtree, B. (1998, March). *Identifying and tracking changing interests*. Retrieved from <http://www.btexact.com/projects/agents.htm>

Soltysiak, S., & Crabtree, B. (1998). Knowing me, knowing you: Practical issues in the personalization of agent technology. *Proceedings of PAAM'98*, London (pp. 467-484).

Zhu, F. M., & Guan, S. U. (2001, October 7-10). Evolving software agents in e-commerce with GP operators and knowledge exchange. *Proceedings of the 2001 IEEE Systems, Man and Cybernetics Conference*, Tucson, AZ (pp. 3297-3302).

ENDNOTES

- ¹ <http://www.infoworld.com/articles/hn/xml/02/03/22/020322hnmcommerce.xml>
- ² <http://www.nngroup.com/reports/wap/>
- ³ <http://www.altavista.com>
- ⁴ <http://www.yahoo.com>
- ⁵ Profile-based text summarization
- ⁶ <http://www.hardwarezone.com>
- ⁷ <http://www.philips.com.sg/news.shtml#5January>
- ⁸ <http://www.compaq.com/products/handhelds/pocketpc/H3870.html>
- ⁹ <http://www.pcworld.com/features/article/0,aid,82004,pg,5,00.asp>

Chapter XXXVIII

Understanding Mobile Web Services (MWS) and Their Role in Integrating Mobile Devices

Samir El-Masri

University of Western Sydney, Australia

ABSTRACT

Web services (WS) have become the industry standard tools for communication between applications running on different platforms, and built using different programming languages. The benefits, including the simplicity of use, that Web services provide to developers and users have ensured integration of Web services architecture by almost all IT vendors in their applications. As expected, with the proliferation of mobile phones, PDAs and other wireless devices, the same requirements of making applications talk across platforms has become necessary on mobile devices. This has led to the mobile Web services (MWS), which are based on the Web services and related technologies like XML, SOAP and WSDL, and which provide the best choice to be used in the architecture for integration of Web services in mobile devices. This chapter discusses WS and MWS in the context of integration architecture, together with their advantages and disadvantages in usage. Since MWS is deployed using wireless technologies and protocols, they are also presented and explained in this chapter.

INTRODUCTION

Web services (WSs) represent the next major chapter of online computing that has enabled seamless integration of application services across the Internet. WSs are the cornerstone towards building a global distributed information system, in which many individual applications will take part. The centre for that global

system will obviously be the WS. As there is no place today for a stand-alone computer, there will be no place for stand-alone applications in the future. Therefore, building a powerful application whose capability is not limited to local resources will unavoidably require interacting with other partner applications through WS. The Internet has been the revolution in networking that links computers and people in a

manner that changes the way we live and work and do business forever. It is believed that WSs will be changing the way we develop software and build applications, in a way that one application will depend on and use many other applications online (El-Masri & Unhelkar, 2005).

The strengths of WSs come from the fact that WSs use XML and related technologies connecting business applications based on various computers and locations with various languages and platforms.

The increase of applications in the mobile world (mobile phones, personal digital assistants (PDAs), etc.) makes it necessary for those applications to communicate with other applications residing on computers on the Internet. WS is a successful architecture for building software applications on the classic network. Recently, Java readied itself for wireless Web services (Yuan & Long, 2002). Microsoft, the leading company in building computer applications, and Vodafone, the leading group in the MS world (Microsoft & Vodafone, 2005a, 2005b) agreed to work together to build standards to facilitate the integration of mobile applications with other applications using a new architecture called mobile Web services (MWSs). Many papers have been published recently in this area (e.g., El-Masri, 2005; El-Masri & Unhelkar, 2005; El-Masri & Suleiman, 2005). The proposed MWSs are to be the base of the communications between the Internet network and wireless devices such as mobile phones, PDAs, and so forth. The integration between wireless device applications with other applications would be a very important step towards global enterprise systems. Similar to WS, MWS is also based on the industry standard language XML and related technology such as SOAP, WSDL, and UDDI. These technologies will be presented with more details in the next sections. Practically, Microsoft .NET framework makes it a simple task to build

a mobile Web application consuming Web services (Arora & Kishore, 2002). Java and IBM also have their own environment for the same purpose.

EXTENSIBLE MARKUP LANGUAGE (XML)

Understanding XML forms the starting point of this discussion, leading into Web services. This is because XML is at the core of WS. XML is a simplified version of SGML (Standard Generalised Markup Language), on which HTML (HyperText Markup Language) is based (Quin, 2004; Ray, 2001). HTML has its own defined tags, which cannot be modified. On the contrary, XML allows users to choose their own tags and elements depending on their need. HTML is a data and presentation language viewed by humans via browsers. XML is a data carrier document and is independent of any presentation using a related technology like XSLT (Extensible Style Sheet Transformation), which in turn is an XML document. The XML document can be viewed with different formats via Internet browsers, PDAs, or mobile phones, or it can be transferred to another XML document. Because of those features, XML represents for IT vendors a brilliant future as a common language and a medium to exchange data independently of the languages and the platforms used by applications. XML is dramatically and rapidly changing technology, and many believe that XML is the next revolution in technology. Below is an example of a simple XML document of a health record of a patient:

```
<?xml version="1.0"?>
<Patient_Health_Record>
  <Patient_Name>      Peter      Lee</
Patient_Name>
```

```
<Patient_Mobile>0405060708</Patient_Mobile>
<Patient_Address>20 Pit Street Sydney
NSW 2000</Patient_Address>
<Patient_Local_Doctor>
  <Doctor_Name> Michelle Fouler</Doctor_Name>
  <Doctor_Mobile> 0415161718</Doctor_Mobile>
</Patient_Local_Doctor>
</Patient_Health_Record>
```

In the example above, there is an element called `<Patient_Health_Record>` that contains four sub-elements: `<Patient_Name>`, `<Patient_Mobile>`, `<Patient_Mobile>`, and `<Patient_Local_Doctor>`. The last sub-element `<Patient_Local_Doctor>` in turn contains two sub-elements: `<Doctor_Name>` and `<Doctor_Mobile>`. Each element or sub-element contains data such as `<Patient_Name> Peter Lee</Patient_Name>` where the data are *Peter Lee*. As shown above, each XML document has its own tags that could be introduced to others by XML Schema (Priscilla, 2002).

WEB SERVICES

The most successful use of XML comes from its role in WS. Earlier technologies like DCOM, CORBA/IIOP, and RMI have been used in integrating applications. These technologies could make two or more applications based on different languages and platforms communicate with one another. The problem is that these technologies are complex and expensive to implement, and they have to be implemented specifically for the applications in question. On the contrary, the WSs that are based on SOAP messaging are simple, easy to deploy, and can be used by any application. WS and SOAP use

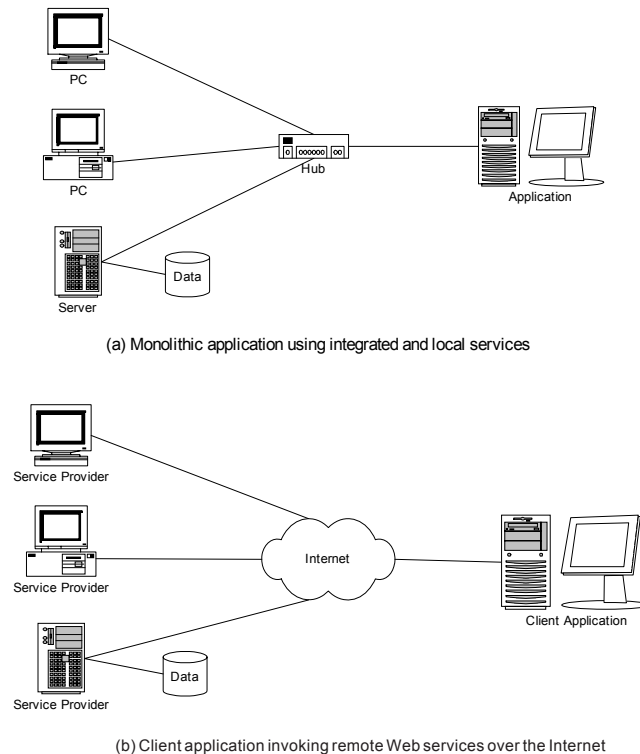
the ubiquitous Internet protocol HTTP as a transport protocol, which makes it even more popular. When an application uses Web services provided on the network, it does not actually integrate into the application computer; it instead invokes or calls from their locations on the network. Once services are called, they will be processed on their own machines, calculate the results, and pass them on to the client application over the Internet. Therefore, in the Web services environment, when we run an application, many machines will be working together to achieve the request^{3/4}that is, the application machine and the services' providers machines (Chatterjee & Webber, 2004)

Web services are resources available on the Internet, while Data Link Library (DLL) is made up of local resources, as shown in Figure 1.

In Figure 1(a), the application uses some DLL and .exe local files, while in (b) the client application uses and consumes some services located on remote machines. The application and service providers' machines connect together via the Internet.

Having mentioned the advantages of WSs, it is fair also to mention their disadvantages. As the Internet is directly involved in WS calls and return results, the performance of the application will depend primarily on the speed of the network. Therefore a slowing network could put the application down. The performance of the application will also depend heavily on the performance of the providers' machines. The consistency of the services provider is also questioned, as sometimes services are not available. Security could also be a concern when hackers may intercept the data on the network. Those disadvantages present real challenges for the future of WS. A lot of research is going to overcome those challenges and make WS a mature and practical technology.

Figure 1. The difference between using integrated and local resources (DLL) (a) and invoking remote Web services over the Internet (b)



UNIVERSAL DISCOVERY AND DESCRIPTION AND INTEGRATION (UDDI)

WSs are usually registered and published through UDDI, which is in turn an XML-based document. UDDI is similar to the well-known yellow pages of the telephone directory. In other words, UDDI represents the yellow pages for WS, where most WS providers register their services. Developers who build applications based on WS resources search UDDI to find suitable and useful capabilities or services represented by functions (methods or behaviours) and classes. UDDI could also provide specifications about the WSs, which can be useful for potential client applications.

WEB SERVICES DESCRIPTION LANGUAGE (WSDL)

WSDL, which could be integrated into or separated from UDDI, provides complete details about WSs in terms of how to use, invoke, and connect to potential client applications.

WSDL describes all the services programmatically available for use, as well as the ports, messages, and protocols. In addition it should contain the XML Schema explaining the structure and the data type of the documents. It specifies in detail the parameters you should send in the message heading for the WSs and the expected response message if any. In fact WSDL describes the structure of the message that should be sent and received by the client

application. WSDL is an XML-based document, which could be generated automatically by some tools like Microsoft .NET framework. Although service providers could communicate with the client applications directly, there is a strong need to register with UDDI which expose them to more customers and make them easier to be found (see Figure 2).

As shown in Figure 2, the application could be a normal Internet client application or a wireless/mobile application. The service provider can also be a mobile provider as well as normal services. Figure 2 shows that an application developer starts the process by searching the UDDI for required services needed in their application. The found methods representing the services must have been registered in advance with UDDI. The second step starts when UDDI provides specifications (WSDL) to the application about how to use and call the services.

SIMPLE OBJECT ACCESS PROTOCOL (SOAP)

Like UDDI and WSDL, SOAP is an XML document that contains an envelope as the root

element representing the message, and which contains an optional header element and a body element. The header is often used for security and encryptions. SOAP is used as a protocol of communications mainly between WS providers and client applications that consume WS. SOAP is carried by HTTP and is firewall friendly (El-Masri & Unhelkar, 2005).

SOAP-RPC

A SOAP message becomes more attractive when it is associated with RPC (remote procedure call), especially to those who suffer from the complexity of CORBA and DCOM. A client application can send a SOAP message containing a call and parameters for an operation (method) that belongs to some service providers. The WS centre or operator passes on that call to the operation or method of a class belonging to the WS provider. The operation returns a response carried back to the client by the WS centre (El-Masri & Unhelkar, 2005).

Figure 2. Discovering WS, developing application through WSDL, and registering by application developer

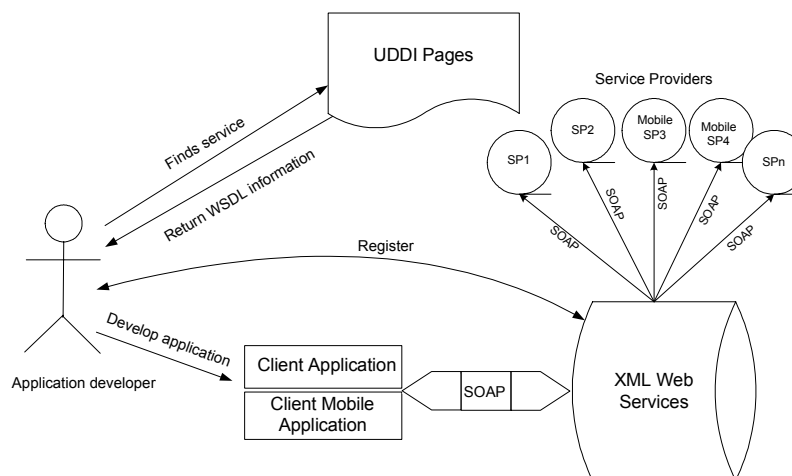
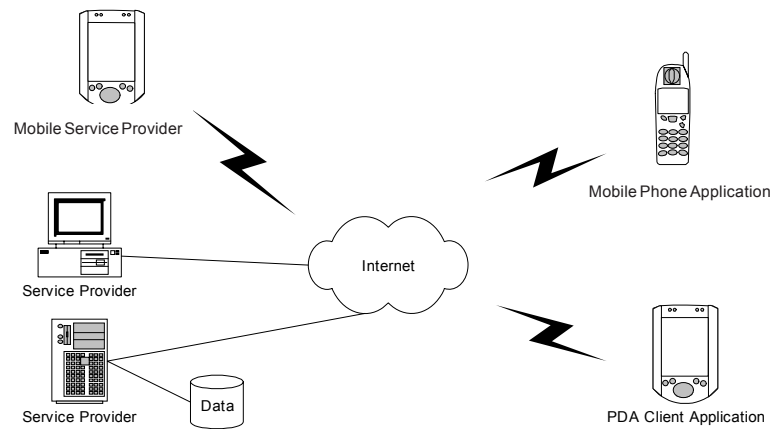


Figure 3. Mobile Web services—client and provider may be mobile devices



MOBILE WEB SERVICES

Let us start with an attempt to understand the meaning of the terms *mobile* and *wireless* devices. Mobile is wireless, but wireless is not necessarily mobile. Wireless devices do not require wires and cables to be connected to a network, and can be static and not moving such as a PC in a Wireless LAN (WLAN), wireless keyboard, and wireless mouse. As for mobile devices, in addition to being wireless, they can move freely in a wider environment and still be connected to the network such as a mobile phone. This chapter deals with the mobile term as it covers the other wireless one.

The scenario where a mobile device such as a mobile phone or PDA can wirelessly access the Internet network, and use and consume Web services, will claim the mobile Web services use. It is also possible that the WS provider would be hosted by mobile devices (El-Masri & Suleiman, 2005) (see Figure 3).

This case exists when a mobile phone can provide its location to some applications, for example parents like to locate their children when they go to school. Another example of a

Web service provided by a mobile phone could be details and information about the owner of the phone for employers so that an employer using a client application can look for the right employees by accessing some services provided on mobile devices (El-Masri & Suleiman, 2005). This will obviously require in both cases the authorisation of the user. It can be concluded that mobile Web services are used whenever there are mobile devices using or providing Web services to consumer applications (Chatterjee & Webber, 2004).

MOBILE WEB SERVICES CHALLENGES

Many constraints make the implementation of WSs in a mobile environment very challenging. The challenge comes from the fact that mobile devices have smaller power and capacities as follows (Chatterjee & Webber, 2004):

- Small power limited to a few hours
- Small memory capacity
- Small processors not big enough to run larger applications

- Small screen size, especially in mobile phones, which requires developing specific Web sites with suitable size
- Small keypad that makes it harder to enter data
- Small hard disk
- The speed of the data communication between the device and the network, and that varies

Having mentioned the current limitations of mobile phones, it is believed that people will use more PDAs than traditional mobile phones, as the former can have a wider screen, a more powerful processor, and larger memory and storage. As for the power and battery, a lot of improvement has been made.

PROXY-BASED MOBILE WEB SERVICES

The most popular MWS is a Proxy-based system where the mobile device connects to the Internet through a proxy server. This system is specifically useful when the mobile device is a mobile phone that has a limited processor capacity, limited memory and battery (see Figure 4).

Most of the processing of the business logic of the mobile application will be performed on the proxy server that transfers the results to the

mobile device that is mainly equipped with a User Interface to display output on its screen. The other important advantage a Proxy server provides in MWS is, instead of connecting the client application residing on the mobile device to many service providers and consuming most of the mobile processor and the bandwidth, the proxy will communicate with service providers, do some processing, and send back only the final result to the mobile device. An example of a client application on a mobile device that needs more than one service provider is a travel planning application, where airlines, hotels, and rental car services are needed at the same time.

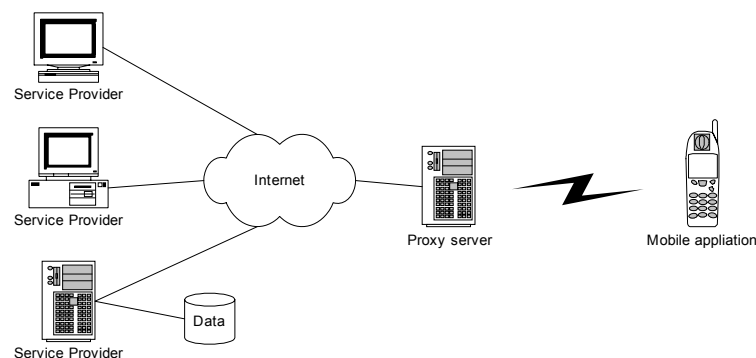
MOBILE DEVICE PROTOCOLS

In the following sections, some mobile device protocols will be explained and discussed to show the infrastructure on which MWSs could be implemented:

TCP/IP

The transmission control protocol/Internet protocol was originally used to transfer data over the Internet network. This platform-independent protocol breaks data into smaller packets, transfers them over the network, and reassembles them back to reproduce the original

Figure 4. Proxy-based mobile Web services



data at the destination computer. TCP/IP is the transfer protocol used to access the Internet and MWSs from PDAs or other devices that support TCP/IP-based network communication (Chatterjee & Webber, 2004).

WAP (GPRS)

Wireless application protocol (WAP) or general packet radio service (GPRS) is used as a transfer protocol for new mobile phones and other WAP-enabled mobile devices to access the Internet and MWSs. WAP has been introduced specifically to access the Internet from mobile phones with limited battery, memory, and processor capacity. In this context, the mobile phone will be used mainly as a user interface (thin client) to access the Internet through WAP Gateway software that deals with encoding and decoding Web sites using XML and Wireless Markup Language (WML) to present data to mobile phones with a text format (Chatterjee & Webber, 2004).

WIRELESS NETWORKING EVOLUTION

MWS operations may be carried out by two different infrastructures. The first is the Wi-Fi wireless environment, which is supported by the increasing number of hot spots implemented by telecommunications companies such as Telestra in large cities to provide access to the Internet for mobile users. Wi-Fi technology offers direct access to the Internet by covering cities with wireless access points connected to the network. The more Wi-Fi hot spots, the better signal and coverage we get.

The other technology is based on the mobile phone network, which is based on cells and satellites. By improving the bandwidth of com-

munication on mobile phones and make it reasonable, the Internet and all its services can be used on that network. In the following sections new technologies will be presented to improve the bandwidth of the cellular network.

GSM

The most popular mobile network is Global System for Mobile (GSM) communications. Within the GSM networking system, each mobile phone is allocated with a single slot of 9.6 Kbps. In new GSMs, this slot has been updated to 14.4 Kbps. It is obvious that this speed is very slow for accessing the Internet from mobile devices no matter what protocols (TCP/IP or WAP) will be used (Mobile Info, 2001).

3G

There is also a new high-speed implementation of GSM called High-Speed Circuit-Switched Data (HSCSD), which can multiply the previous speed (14.4 Kbps) by 4 to reach a transfer rate up to 57.6 Kbps. This is the 3G or *third-generation* emerging technology of mobile networking with a rate a bit higher than dial-up Internet access. 3G technology is increasing, but it is not yet fully implemented (Mobile Info, 2001).

4G

The growing need to access the Internet from mobile devices pushes networking developers to develop faster mobile networking with an aim to reach a broadband speed over a mobile network. A mobile broadband speed can be reached by *fourth generation* (4G). 4G is under test and will be implemented by different

technologies such as FLASH-OFDM for Nextel's iDEN, EDGE for GSM, and EV-DO for CDMA (Mobile Info, 2001).

CONCLUSION

In this chapter, the principle of Web services was addressed, along with related technologies such as XML, SOAP, WSDL, and UDDI, and the great benefits they have brought to the development and integration of software applications over the Internet. The extension of Web services into mobile Web services, where mobile devices can provide services to other applications and can consume services from other computers on the Internet network, was also demonstrated. To effectively access the Internet and Web services from mobile applications built on mobile devices, there was a need for a larger bandwidth on the mobile wireless network. It was shown that larger bandwidth is already available using 3G and 4G technology. The new Wi-Fi technology has also been explained as an extension to the Internet network. The main challenges to the mobile Web services would be the consistency of the services, their availability, and the coverage of the mobile network as well as the security issues, as the wireless environment is even more vulnerable.

REFERENCES

- Arora, G., & Kishore, S. (2002). *XML Web services—professional projects*. Cincinnati, OH: Premier Press.
- Chatterjee, S., & Webber, J. (2004). *Developing enterprise Web services*. Englewood Cliffs, NJ: Prentice-Hall.
- El-Masri, S. (2005). Mobile comprehensive emergency system. In *Proceedings of the 2nd International Conference on Innovations in Information Technology*, Dubai, UAE.
- El-Masri, S., & Suleiman, B. (2005). Providing Web services on mobile devices. In *Proceedings of the 2nd International Conference on Innovations in Information Technology*, Dubai, UAE.
- El-Masri, S., & Unhelkar, B. (2005). Modelling XML and Web services messages with UML. In *Proceedings of the Information Resources Management Association 16th International Conference*, San Diego, CA.
- Microsoft & Vodafone. (2005a). *Convergence of PC and mobile applications and services*. Retrieved February 10, 2005, from http://www.microsoft.com/serviceproviders/mobilewebservices/mws_whitepaper.asp
- Microsoft & Vodafone. (2005b). *Mobile Web services technical roadmap*. Retrieved February 10, 2005, from http://www.microsoft.com/serviceproviders/mobilewebservices/mws_tech_roadmap.asp
- Mobile Info. (2001). *4G—beyond 2.5G and 3G wireless networks*. Retrieved August 8, 2005, from <http://www.mobileinfo.com/3G/4GVision&Technologies.htm>
- Priscilla, W. (2002). *Definitive XML Schema*. Englewood Cliffs, NJ: Prentice-Hall.
- Quin, L. (2004). *Extensible Markup Language (XML)*. Retrieved July 10, 2004, from <http://www.w3.org/XML/>
- Ray, E. T. (2001). *Learning XML*. Sebastopol, CA: O'Reilly & Associates.
- Yuan, M. J., & Long, J. (2002). *Java readies itself for wireless Web services*. Retrieved January 20, 2005, from <http://www.javaworld.com/javaworld/jw-06-2002/jw-0621-wireless.html>

Chapter XXXIX

Push–Multicasting to Wireless Devices Using Publish/Subscribe Model

Jon Tong-Seng Quah

Nanyang Technological University, Singapore

Chye-Huang Leow

Singapore Polytechnic, Singapore

ABSTRACT

Push technology is a kind of technology that automates the information delivery process without requiring users to request the information that they need. Wireless has experienced explosive growth in recent years; “push” will be the predominant wireless service delivery paradigm of the future. We can expect a large number and a wide variety of services, alerts and messages, such as promotional content, to be delivered to consumer’s phones or PDAs. To push information to wireless device becomes a challenge because of the problem of intermittent communication links and resource constraint on wireless devices as well as limited bandwidth. This chapter describes an efficient multicasting mechanism that “pushes” pre-specified information to groups of wireless devices. The mechanism is able to operate with limited bandwidth and also overcome the connectivity problem. A framework has been designed that implements the concept of push technology to multicast the sales information via wireless technology. The design of a message-oriented system for wireless information is described and that is followed by the implementation details that are compliance to Java Message Service (JMS).

INTRODUCTION

Today, the vast scale and scope of current online information sources on the Internet makes it difficult to find and process relevant informa-

tion. Finding a specific piece of information on the Internet requires time-consuming search. Hence, automation to push pre-specified information to the user seems like the next logical step to solve these problems.

We are moving towards third-generation wireless technology where multimedia applications are supported in wireless handheld devices and hand phones. It is believed that push will be the predominant delivery methodology in wireless device services. This is due to the problem of servers being unable to push data to clients who are disconnected. However this is not an issue in GPRS wireless networks. In GPRS wireless networks, the users will be always connected to the Internet. Hence, we can expect a large number and a variety of services, alerts, and messages, such as promotional content, to be delivered to consumers' mobile phones or PDAs in real time. In addition, there are some constraints in wireless technology, such as the small memory capacity in the devices, limited bandwidth, and the high cost of information searching on the wireless network. Push information to wireless handheld devices will save a great deal of time and money compared to surfing the Internet via WAP technology. Thus, this brings forth the idea, to create a "wireless push" channel to push information to wireless devices in real time.

Furthermore, from a software point of view, in order to cope with the limited bandwidth problem, this research will study an efficient multicasting mechanism to push sales information to a group of members over the wireless network. A simple application and framework about the Internet selling process has been designed, whereby the sales information like product catalog will be multicast to the interested mobile users in real time. The Publish/Subscribe messaging model has been used as a message delivery mechanism in this framework because of the capability of the messaging system to multicast information to a specific group of recipients. The following three criteria must be satisfied in order for a piece of information to be considered suitable for delivery using push mechanism:

- The kind of information desired must be known ahead of time like stock quotes, new headline, and so forth.
- Searching for such information must be an inefficient use of the user's time.
- The user must want this information regularly.

DATA DELIVERY MECHANISM

The paper "Data in Your Face: Push Technology in Perspective" (Franklin & Zdonik, 1998) presents some ideas on data dissemination in order to provide a framework for thinking about push technology. The authors have outlined several options for data delivery, and the comparison of their characteristics is illustrated below:

- Client Pull vs. Server Push—Pull based is a request-response operation, which is client-initiated transfer of information from server to client, whereas in push-based operation, the server initiates the transfer.
- Aperiodic vs. Periodic—Aperiodic is an event-driven operation, where the transfer of information is triggered by an event. In periodic delivery, the transfer information is performed according to a pre-arranged schedule.
- Unicast vs. 1-to-N—With unicast data communication, the data to be transferred is sent from one source to one destination, while multicast and broadcast are 1-to-N data communication. In multicast data delivery mechanism, the transfer data is sent to specific subsets of clients, whereas in broadcast data delivery mechanism, the transfer data is sent to an unidentified set of clients that can listen to it.

Periodic push has been used for data dissemination in many systems, for example an Internet mailing list that regularly sends out mail

to users. Aperiodic push, which is based on Publish/Subscribe protocol, is becoming a popular way to disseminate information to end users.

Benefits of Push Technology

The paper “‘Push’ Technologies: Reborn for Business” (Farber, 2000) summarized the benefit of push technology into one sentence:

It’s mirroring what a really good human assistant would do if all they had to do was sit around and watch out for you.

Push technology has advantages for both end users and content providers. For end users, it significantly improves the response time of accessing Web content. Clients do not have to waste cycles and network traffic to poll servers.

For content providers, they are allowed to target their audiences in a more direct way, which results in a cleaner business model. Servers can use more processor time for data production rather than to process numerous client requests and send much data over the network. Furthermore, servers can better manage the amount of data transferred over the network as it delivered useful and interesting information to clients.

Constraint and Challenges in Push Technology

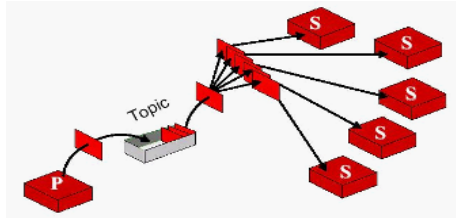
“Data in Your Face: Push Technology in Perspective” (Franklin & Zdonik, 1998) has identified some issues regarding push technology. Current Web casting applies pull technology instead of push due to limitation of HTTP protocol. HTTP protocol is a kind of pull-based protocol. Broadcasting and multicasting are not widely used as it potentially causes bandwidth problems if multi-clients request the same data through unicast data communication. In addition, the article “What’s All Wrong with Today’s Push Technology?” (Berst, 1997) has pointed

out that push technology should be built on top of multicast data communication. It is not advisable to implement push technology on top of unicasting data communication due to inefficiency that will result as the user based grows. Furthermore, the article “The Push Technology Rage...So What’s Next” (Gerwig, 1997) also pointed out that the next step of push technology is a multicasting delivery mechanism that will enable true push technology to millions of users across the Internet. The Internet still uses the basic one-to-one ratio of one request for information sent to one computer at a time. Therefore Internet service providers (ISPs) will have to update their networks to handle multicasting, which will enable content providers to “broadcast” data to a large number of users rather than sending content using a one-to-one model. Furthermore, this article highlighted a need for a standard message in pushing technology as a common way to tell users what is in a content channel and what the users need to view it. XML as a standard of W3C has been briefly discussed in this article.

The article “When Push Comes to Shove: Push Technology Is All the Rage—What Does This Mean for Java?” (Blundon, 1997) discussed “Web casting” or “push” technology and how it can dramatically increase the productivity of the Internet or intranet in data communication. In short, the main challenges in push technology are:

- Multicast or broadcast information to those users who are interested to prevent clogging the network traffic.
- Standardize the push message to all channels, especially to push information to different platforms or devices, like various manufacturers’ wireless phone, Palm OS, IPAQ, Visor, and so forth.
- Security will be an issue because the user allows information or software components to be downloaded to his local devices.

Figure 1. Publish-and-Subscribe messaging



PUBLISH-AND-SUBSCRIBE MESSAGING—FRAMEWORK FOR PUSH SELLING IN WIRELESS NETWORK

The Pub/Sub messaging model is where there is only one sender and one or more receivers. Rather than queuing for sending and receiving messages, there is a topic, which is equivalent to an individual newsgroup. Interested clients subscribe to receive all messages sent to those topics, and they can publish to those topics if they wish to. The number of publishers and subscribers grows and shrinks over time. Figure 1 illustrates the process of Pub/Sub messaging where “P” is publisher and “S” is subscriber. This model is suitable for the push model application, as consumers are delivered messages without having to request them.

Wireless JMS

JMS provides asynchronous and message-based transportation. Using message queues hosted on both the client and the server side, JMS applications can be operated in disconnected mode, and data synchronization occurs transparently and immediately without user intervention. In addition, it provides an ideal abstraction layer for developing mobile applications and also increases the scalability of the system. Many mobile devices can simultaneously send messages to a server. When messages arrive at the server, they are added to an inbound queue

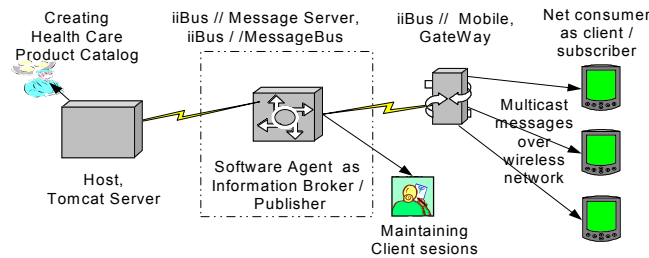
and can be dealt with when resources are available, or can be forwarded to other servers for load sharing. In practice, JMS allows wireless services to operate more responsively, to recover from sporadic network outages easily, and to allow mobile applications to be operated off-line. Finally, JMS allows messaging to be implemented elegantly atop Bluetooth (Salonidis, Bhagwat, Tassiulas, & LaMaire, 2005), Wireless LAN (Wu, Long, & Cheng, 2002), GPRS (Vergados, Gizelis, & Vergados, 2004), UMTS (Yang & Lin, 2005), and Mobitex (Wei & Jost, 2003).

Mobility Issues

In this section we will look at the mobility issues of developing mobile applications. Mobile applications present many challenges for software developers and require them to deal with the issues not present in wired line systems. The following summarizes the issues:

- **Intermittent Communication Links:** Mobile devices often lose network coverage. To deal with these intermittent communication links, the mobility application should provide disconnected operation and guarantee important data reaching mobile devices.
- **Resource Constraint:** Mobile applications must be optimized aggressively for small ROM and RAM footprints, as well as for low usage of CPU cycles and battery power.
- **Multiple Bearers:** Currently there are many different bearers in wireless networks like SMS, GPRS, Infrared, Bluetooth, and HTTP. Developing one mobile application and running it in different bearers becomes the challenge of developers.
- **Multiple Platforms:** There are various platforms in the mobile communicator

Figure 2. Architecture design of Push Selling—Multicast Message to Wireless Devices



device market such as Palm OS, Pocket PC, Symbian, and Windows CE. It is advantageous to develop applications in such a way that they can be run on various platforms, without requiring substantial modifications.

Building mobile applications using a middleware with JMS implementation can solve most of these issues and problems. iBus//Mobile JMS middleware has been selected to prove the concept in this chapter. The following sections will elaborate on the implementation details of the application model proposed here.

Architecture Design

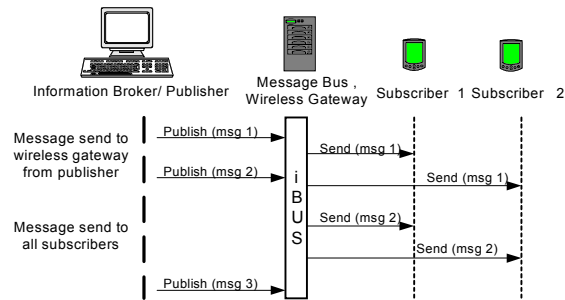
A simple application and framework has been designed as shown in the Figure 2. This push-based application implements the Internet selling process whereby the product catalog created by the seller will be multicast to mobile users. The proposed model mainly comprises three types of nodes:

1. The *host* provides the base data that is to be disseminated. In our example system, we assume the core business of the system is to sell health care products and services. The system will disseminate the products' information in the form of a catalog. The system provides a GUI for

creating the product catalog. This GUI is a JSP (Java Server Page) interface running on top of a Tomcat servlet engine. This GUI enables the user to create the product catalog, which will then be sent by the *host* to an agent known as an "information broker."

2. The *information broker* plays a vital role in acquiring information about the product catalog, and adds on additional value or data to that information and then pushes this information to clients. The information broker is running on top of iBus//MessageServer and iBus//MessageBus. In the event that this broker is being triggered, it will create a topic and all users who are subscribing to this topic should be able to receive the messages. This broker behaves like a "publisher" to multicast the product catalog to those active subscribers. The data delivery mechanism is based on the Publish/Subscribe model, which will deliver the information to a specified set of interested clients. In addition, the clients' sessions and configuration data will be kept and maintained in this node (i.e., the information broker server).
3. *Clients*, which are net consumers, will receive such selling information without requesting it via iBus//Mobile. iBus//Mobile provides a gateway for transferring

Figure 3. Publish/Subscribe Messaging in 1-N push technology



the messages to wireless devices. From the JMS provider's point of view, the gateway is a regular JMS client. From the mobile device's point of view, the gateway is a communications hub and message format translator. A simple MIDLET application has been developed using Light-Weight JMS Client Library. This library is provided by iBus/Mobile. The MIDLET will be loaded into a client's PDA and runs as a client's agent to receive the incoming push information. The MIDLET is built on top of a PALM operating system.

Message Delivery Mechanism—Publish/Subscribe Model

The message delivery mechanism implemented in this system is based on the Publish/Subscribe model, illustrated in Figure 3. The transfer data is pushed to specific subsets of clients. The following describes the message flows:

- Once a seller has created a product catalog, the system will trigger a Servlet to send this product catalog in the form of message to the information broker.
- The broker will create a topic named "push_info." In order to receive this mes-

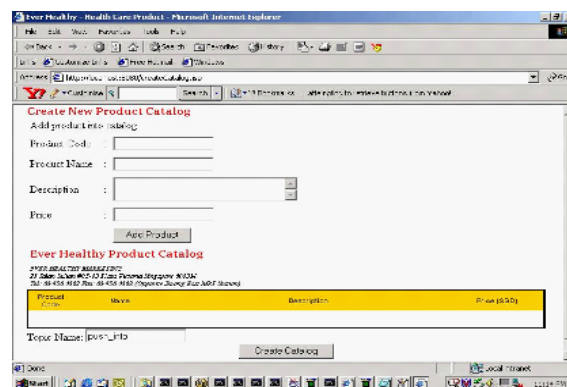
sage, the mobile users need to subscribe to this topic.

- The broker will start publishing its first message, and iBus middleware will start sending the message to all subscribers. The middleware ensures that all subscribers receive the message of this topic "push_info."
- The broker is allowed to publish a second message, although the first message has not completely been sent to all subscribers.
- As time passes, new topics may be created by the broker. Users can log on using their user identification registered with the broker to browse through the Web pages displaying available topics and select/unselect topics according to their interests.

Result

This section presents the results of the implementation as a proof of concept. Figure 4 shows a user interface for sellers to create their product catalog. The GUI is developed in JSP and runs in Tomcat Servlet Engine. The page consists of several input textboxes for users to key in catalog information like product code,

Figure 4. User interface for creating product catalog



product name, product description, and price. The seller is allowed to create multiple items per catalog by clicking on the “Add Product” button. The seller publishes the catalog under a specific topic by keying in the topic name at the bottom of the page. Once “Create Catalog” button has been clicked, a product catalog will be created and the page will trigger a Servlet known as “triggerPushAgent.” The Servlet writes the catalog message to a file, named “productcatalog.txt,” and then it will instantiate the Information Broker. The broker will read the message from that file and multicast the message to all subscribers. The subscriber needs to enter the topic name to which he wishes to subscribe. His device will then receive all messages under this topic.

Once the subscriber has logged on, the application will be in listening mode and wait to receive incoming messages.

Figure 5 shows that the two subscribers, User001 and User002, have received the product catalog.

CONCLUSION AND FUTURE DIRECTION

This push-based selling system enables mobile users to receive product information in real time. The crucial element of the system is that the sender assigns messages to topics and not to a particular remote object, and then the receivers couple the messages based on the topics. The system provides a company as content provider to simultaneously push its selling information to multiple handheld devices such as PDAs. Therefore, from the perspective of the sender, it is more effective and efficient in transmitting large amounts of data to a group of receivers in real time. No additional user interfaces or efforts are required in selecting or customizing the group of receivers. In other words, subscribers received the selec-

Figure 5. Clients' application—receiving and displaying product catalog



tively streamed information based on their indicated interests—as and when contents are added to topics at the broker, once connectivity can be established while they are on the move.

Furthermore, from the sellers' perspective, it allows targeting of their audiences in a more direct way, which results in a cleaner business model. The server can better manage the amount of data transferred over the network as it delivers information on topics that clients have subscribed to. This significantly enhances scalability and uses the available network bandwidth with maximum efficiency.

From mobile users' point of view, they do not need to spend a lot of time and money in order to get the information they want. The information is automatically pushed into their wireless handheld devices. We are currently working with two groups of early adopters of the technology. The first group consists of investors who want real-time feeding of stocks information into their devices for decision making. The second group consists of medical practitioners who need to get the latest condition updates of their inpatients and use the information to decide whether urgent attention is needed by those patients under their care. Both groups of users have given their thumbs-up for the trial system.

We are working on ways to better the current implementation. Research on IP multicast for mobile hosts, especially in a wireless environment like Mobile IP, is recommended to enhance the system and to realize true multicasting, which is not just from a software perspective, but also from a network infrastructure point of view. A multicast agent scheme for mobility support of IP multicast in the Internet is possible. The approach is to use three-layer architecture for multicasting to mobile hosts, and introduce multicast agents that serve as access points to the multicast backbone by mobile hosts and are located close to the current locations of mobile group members. At the Internet level, multicast agents are simply multicast routers participating in multicast routing.

Besides the multicast delivery mechanism, it is useful to also study the broadcast delivery mechanism. Broadcast is considered a type of push-based data delivery mechanism. The basic idea is, the server periodically broadcasts the sales information, while client agents monitor the broadcasted messages and only receive messages that they require. The advantage of this approach is that other clients who are also monitoring the broadcast do not directly affect the performance of any other client receiving data from the broadcast.

Furthermore, we are also researching the use of mobile agents in wireless handheld devices because they will overcome wireless network limitations such as low bandwidth and intermittence disconnection problems. Agents can act as buyers and sellers, and communicate and negotiate with each other autonomously (Schmid & Quah, 2004; Quah & Seet, 2004; Quah, Leow, & Soh, 2004).

REFERENCES

- Berst, J. (1997). *What's all wrong with today's push technology?* Retrieved from <http://www5.zdnet.com/anchordesk/story/>
- Blundon, W. (1999). *When push comes to shove: Push technology is all the rage—what does this mean for Java?* Retrieved from <http://www.javaworld.com/javaworld/jw-04-1997/jw-04-blundon.html>
- Farber, D. (2000). *Push technologies: Reborn for business*. Vineyard Haven, MA: Vineyardsoft Corporation.
- Franklin, M., & Zdonik, S. (1998). Data in your face: Push technology in perspective. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Seattle, WA (pp. 516-519).
- Gerwig, K. (1997). The push technology rage...so what's next? *ACM Digital Library, The Craft of Network Computing*, 1(2) 13-17.
- Quah, J. T. S., Leow, W. C. H., & Soh, Y. K. (2004, June 21-24). Mobile agent based e-learning system. In *Proceedings of the 2004 International Conference on Information and Knowledge Engineering*, Las Vegas, NV (pp. 256-265).
- Quah, J. T. S., & Seet, V. L. H. (2004, June 10-13). Improving usability of WAP portal through adaptation. In *Proceedings of the 7th International Conference on E-Commerce Research*, Dallas, TX (pp. 407-420).
- Salonidis, T., Bhagwat, P., Tassioulas, L., & LaMaire, R. (2005). Distributed topology construction of Bluetooth wireless personal area networks. *IEEE Journal on Selected Areas in Communications*, 23(3), 633-643.

- Schmid, A., & Quah, T.-S. (2004). Synergetic integration of aglets and e-speak in e-commerce. *Informatica—An International Journal of Computing and Informatics*, 27(4), 391-398.
- Vergados, D. D., Gizelis, C., & Vergados, D. J. (2004). The 3G wireless technology in tactical communication networks. *IEEE Vehicular Technology Conference*, 60(7), 4883-4887.
- Wei, S., & Jost, A. G. (2003). Virtual socket architecture for Internet access using Mobitex. In *Proceedings of the IASTED International Conference on Wireless and Optical Communications* (Vol. 3, pp. 549-554).
- Wu, H. T., Long, K., & Cheng, S. D. (2002, June). Performance of reliable transport protocol over IEEE 802.11 wireless LAN: Analysis and enhancement. In *Proceedings of IEEE INFOCOM 2002* (pp. 599-607).
- Yang, S. R., & Lin, Y. B. (2005). Modeling UMTS discontinuous reception mechanism. *IEEE Transactions on Wireless Communications*, 4(1), 312-319.

Section VIII

Method

Chapter XL

A Methodology for M–Transformation of Small and Medium Enterprises (SMEs) and its Application in Practice Using CBEADS©

Ioakim (Makis) Marmaridis
University of Western Sydney, Australia

ABSTRACT

Organisation worldwide come to realise that in the ever changing business world, survival and success is closely linked to adopting information and communication technologies (ICTs). Along with the technology however, organisations have to also adjust their processes to take full advantage of the potential ICTs have to offer. This process, of technology adoption, linked with with process adjustment and re-engineering is called e-transformation. For organisations that have successfully e-transformed, it is now necessary to become more agile through the adoption of mobile technologies. This adoption leads to the need m-transformation which is the next logical step from e-transformation. In this chapter we define m-transformation and present a methodology that SMEs can adopt in order to m-transform. The methodology takes into account the special characteristics SMEs have and allows them to leverage their strengths towards a smoother m-transformation process. Furthermore, we show how m-transformation can be practically applied, and in doing so we introduce our technology platform called CBEADS. Finally, we present some of the lessons learnt and demonstrate how SMEs may progress through the adoption of mobile technologies into their operation into gaining increased competitiveness and a global reach.

INTRODUCTION

Mobile technologies are a key influence in any attempt at globalization of business (Unhelkar,

2004). Therefore, what was once understood under the banner of “e-tranformations” (or electronic transformations of organisations) now needs to be understood within the context of

mobility. This leads to the idea of mobile-transformation or “m-transformation.” M-transformation is the process of transitioning an existing organisation into the mobile business world. The earlier e-transformations capitalized on the connectivity accorded by the ubiquitous Internet (Arunatileka & Ginige, 2003; Ginige, 2002). M-transformation is the next logical step for these organisations.

However, the many diverse benefits as well as the challenges of m-transformation are not well known, especially in the SME sector. As a result, there is uncertainty and trepidation. Furthermore, the mobile technology itself is in the process of development and is not fully matured. Based on our spot interviews with some of the SMEs that we had earlier e-transformed, we discovered that the aforementioned reasons were contributing to their reluctance to take up m-transformation. This has led to our interests and desire to investigate further this challenge and discover a set of simple yet consistent requirements leading to a methodology that can be applied in m-transformation. The scope of our study is mainly focused right now on SMEs, given their unique characteristics as well as our knowledge base of working with them in the past (Arunatileka & Ginige, 2003; Ginige et al., 2001; Marmaridis, Ginige, & Ginige, 2004a; Marmaridis, Ginige, Ginige, & Arunatilaka, 2004b; Kazanis, 2003).

This chapter starts with our working definition of m-transformation. It then explores some of the benefits that are due as a result of m-transformation. This is followed by a discussion on how SMEs perceive m-transformation—this section is meant to help us uncover and investigate potential disconnects between our understanding of mobile benefits and that of SME users. Next, we almost exclusively put the focus on the unique characteristics SMEs present and how they affect the uptake of m-transformation in the context of currently available mobile technologies and corresponding

business process mainstream thinking. With a solid understanding of the unique SME characteristics, we then proceed to describe the requirements for each aspect of m-transformation, namely ICT infrastructure, business process adoption, and m-transformation methodology. Finally, we describe the methodology in detail and provide insight from our experiences in applying this methodology to a number of SMEs wishing to m-transform their business.

BACKGROUND TO THIS WORK

Before diving further into the details of SME m-transformation as they are expressed via requirements for IT systems and technologies, it would be wise to offer some background information about the projects that we have so far undertaken and some of the technology used in those, namely CBEADS©. A number of years we identified the need for SME organisations in the greater Western Sydney region of NSW, Australia, to increase their uptake of technology in order to become competitive in the global economy and survive. A vehicle for us achieving this was through the process of e-transformation, where SMEs are encouraged and assisted in their uptake of computers and other information and communication technologies (ICTs) to better their business overall.

In order for us to facilitate the uptake of ICTs, we have also developed a framework known as CBEADS©, which stands for Component-Based E-Application Development (and deployment) Shell and provides a low overhead, low-cost infrastructure for e-business applications development and deployment. Along with the technology we also developed a comprehensive methodology for e-transformation, a concise map of which is our e-transformation roadmap that offers step-by-step guidance to organisations embracing e-transformation.

Having successfully applied the methodology and the CBEADS© technology framework to e-transform a number of SMEs, we are now seeing the need to assist those same organisations move into the next step of m-transformation. The overall approach to doing this is similar to that of e-transformation where we start with a methodology to m-transformation backed by ICTs and tools to achieve those goals faster. We include in this chapter our work in the area of deriving the methodology for m-transformation so that it can meet the unique requirements for SME organisations wishing to undertake an m-transformation journey. A lot of the findings and ideas presented here we strongly believe apply universally to SMEs elsewhere in Australia and the rest of the world. With the hope that the information we provide here will be useful to you, the reader, let us proceed with painting a picture of what SME m-transformation looks like.

DEFINING AND UNDERSTANDING M-TRANSFORMATION

The area of m-transformation is actively researched (Unhelkar, 2004), while on the other hand there are many m-enabled software products that try to capitalize on the high growth and demand for this area in businesses. Consequently, the process and scope of m-transformation remains undefined and not properly understood. For the purposes of this chapter, we shall use our working definition of m-transformation which is consistent with our view of it as the next logical extension to e-transformation.

We define m-transformation as “the evolution of business practices via the adoption of suitable processes and technologies that enable mobility and pervasiveness.” In accordance with this definition, we see the adoption of m-

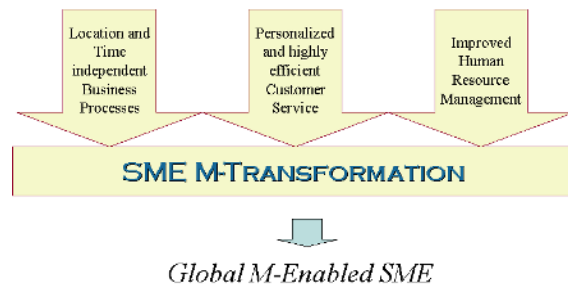
transformation as having two sets of prerequisites: those of business process adoption and those of adoption of suitable mobile technologies and products. Without proper understanding of both these aspects of mobile transformation, embarking on an m-transformation trip can result in a reduced set of benefits flowing back into the organisation or, at worst, failure of the transformation process.

Despite its challenges, though, it is essential that m-transformation is seriously considered by organisations. This is because the advantages of transitioning to a mobile business are also not unfounded. Some of the main benefits that we see stemming from a successful m-transformation are as follows:

- Better information flow between systems, since decisions can be taken quicker by making the necessary information available to the key decision makers faster.
- Improved customer projected image, since the adoption of mobile business processes enables the business to serve the customer in a personalized manner—adding to the projection of a forward-thinking and dynamic image of the company.
- Rapid and dynamic customization of products and services being offered by the organisation—depending on the location and density of potential customers at a particular time and place.
- Increased internal efficiency in managing human resources. This primarily results from the fact that in an m-organisation, information, and people are not being desk-bound anymore.
- Improvement in employee morale coming from flexibility in the workplace—especially with the possibility of telework (Ranjbar & Unhelkar, 2003).

The aforementioned benefits can be summarized and grouped into three major areas

Figure 1. Major benefits from m-transformation of SMEs



positively influencing m-transformation, as shown in Figure 1 (Marmaridis & Unhelkar, 2005).

With these major benefits—both tangible and intangible—that m-transformation promises, it comes as no surprise that there are a large number of SMEs—especially in the Western Sydney region that the authors have worked with—wishing to m-transform themselves. This may well extend the reach of these SMEs to global markets, as envisaged in Figure 1. However, as against a typical medium to large organisation (like a bank, an insurance company, or an airline), SMEs have some special characteristics and present special needs that can be an impediment to their entry into the mobile business arena.

The section that follows describes in more detail what those characteristics of the SMEs are and how they could be overcome. However, let us start with putting m-transformation in the context of the SMEs.

M-TRANSFORMATION AND THE SME

Irrespective of our academic definitions of m-transformation and our view of its current status and future directions, the actual understanding of what it means to the SMEs has a

major influence on its success or failure. The end recipients of our work can have a different view and interpretation of mobile transformation than the ones we formulate. From our experience, this has certainly been the case with SMEs and their view of m-transformation in general. Therefore, further discussions and explanations beyond formal definitions are required in this study.

SME Interpretation of M-Transformation

Whereas we, the researchers, saw opportunities for growth and a host of benefits stemming from m-transformation, some of the SMEs we worked with saw disturbing changes to their business and their ways of doing things associated with high costs of implementation. They could certainly appreciate that having the abilities to remotely access information could be, at times, invaluable. However, SMEs are not ICT experts, nor do they wish to understand the intricacies of mobile technologies. They want something that is simple and easy to use, and just works. Talking to SMEs that are fixated in this mindset about moving into m-transformation was a challenge. It was obvious that we needed to understand the real cause that established this mindset in the first place and then the reasons that were fuelling it. The answers to this came into two parts. Firstly, SMEs have a unique “results-based” view of doing business. The SME characteristics are unique compared to larger enterprises, particularly due to significant lacunae in strategic planning. Secondly, their understanding of ICT technologies that could be suitable for mobility is very fragmented and incomplete; what they seemed to be missing the most was the reassurance that those technologies once implemented will work with each other and be consistently and easily accessible.

Table 1. SME characteristics affecting m-transformation

SME Aspect	SME Characteristic
Staff	Small headcount, affected by seasonal conditions. For example, departure of an IT staff member or power user may mean the entire IT support has vanished. This can create challenges if mobile transformation depends on a particular staff member.
	Limited time and critical availability of business staff. Training users can prove a significant challenge.
	Lack of ICT knowledge and lack of time and interest in gaining that knowledge.
Clients	Each client is highly valued. Client primarily dictates the business. Views of clients on mobile business processes critical in attempting m-transformation.
ICT Systems	Very small budget available covering mostly ongoing running and maintenance with no provisions for new developments.
	Lack of in-house ICT expertise and minimal development in-house.
	Ad-hoc systems establishment and lack of any strategic planning for ICT. Mostly off-the-shelf, point solutions used, making mobile systems integration a challenge.
Decision Making	Decisions are quick, usually taken by the owner/proprietor, and the conditions for decision making are very volatile. Individual personalities may prop in the m-transformation decision-making process.
	Organisational knowledge is tacit and centralized in few key staff. Concern for “loss of turf” or area of influence due to m-transformation is significant.

SPECIAL CHARACTERISTICS OF SMES

Let us consider the special characteristics of SMEs that influence their m-transformations. We were able to glean these characteristics through closely working with numerous SMEs in the corresponding e-transformation processes mentioned at the start of this chapter. Table 1 illustrates those in relation to staff, clients, IT systems, and decision-making needs.

Let us now examine how each of these impact on the SME operations and could therefore influence m-transformation.

Small Staff Count Affected by Seasonal Conditions

SMEs come in all sizes, from one-man operations to over 200 staff. The typical SME size however is between 5 and 100 people. Two distinct characteristics exist in relation to the SME staff, the fluctuation in numbers and

casual nature of work, as well as the over-reliance on particular staff. SME staff numbers greatly fluctuate in response to seasonal factors and varying demand for products and services. Therefore most staff tend to be employed on casual or short-term contract basis. There are only a few people within each company that have a sense of continuation with it.

Due to the few permanent staff, there tends to also be over-reliance on some of those. For instance, the departure of an IT staff member or a computer power user may significantly drain the IT expertise of a particular company. This can lead to challenges in the mobile transformation, as it may be heavily dependant on particular staff members within the organisation.

Small to Non-Existent ICT Budget

The typical budget for an SME of average size is \$3,000 AUD or less per annum, based on our experience. This amount, which for a large enterprise might equate to just three days of

onsite work for an external consultant, will allow an SME to refresh some of their aging hardware for instance. This budget is typically aimed at ongoing running and maintenance of the ICT infrastructure, as opposed to development or acquisition of new technologies for m-transformation.

Lack of Permanent In-House ICT Expertise

In-house ICT expertise is very scarce in SMEs. Organisations of 100 people and under typically have no in-house IT staff due to cost implications. Instead, they rely on external IT contractors to provide, install, and maintain equipment for them as required.

In our experience, reliance on IT contractors and lack on in-house IT staff can significantly hinder the ICT infrastructure planning for m-transformation. IT contractors are only called in when there is a problem requiring their attention, and they tend to operate in “fire fighting” mode. As a result, there is little done in planning of technologies for mobility and other areas.

Staff Have Limited Time and Critical Availability

Because of the small headcount in SMEs, roles tend to be broad, with no backup people to fulfil those should the primary person become unavailable. Also most staff are occupied with carrying out important, day-to-day business functions. These two facts pose significant challenges to the uptake of m-transformation since staff training can prove a challenge. Unlike a large corporation, most SMEs cannot afford to have some of their business staff taken out of their productive capacity for days at a time. A company might significantly suffer, for instance, if their accountant was being

trained in a new IT system for two days in a row. A lot of the accounting work would either lag behind or not get done at all.

ICT Systems Establishment is Done Ad-Hoc

There is a significant lacunae in strategic planning in SMEs, especially in the area of ICT systems to aid m-transformation. Most ICT systems in use by SMEs tend to be off-the-shelf, point solutions with very poor if any integration with each other.

The rapidly changing business environment requires SMEs to select and implement ICT solutions fast and with minimum cost. Coupled with their lack of expertise to forward plan those solutions, they tend to rely on commodity hardware and package-based software for particular functions.

As a result, in a typical SME one can find a host of assorted applications, with no standards in regards to the hardware used, no standard operating environment (SOE), or any thoughts of interoperability provided between applications or data stores.

In such an environment, mobile systems’ integration can prove to be a challenge, especially given the resources SMEs have available towards such integration work.

Organisational Knowledge is Centralized and Tacit

The organisational knowledge in an SME tends to be tacit and centralized in a few key staff. It is not uncommon in our experience that only a single person in the company (typically the owner) has full knowledge about the company’s production, sales, financial, and client information. In larger SMEs this key knowledge might reside with two to three staff in total.

With the knowledge being so centralized, there are obvious and significant concerns for “loss of turf” or area of influence due to m-transformation. The few key staff members that hold a lot of influence in the organisation may view m-transformation as taking away from their influence and power. It is therefore possible for some to oppose the move for m-transformation, significantly hindering its progress and chances of success.

Very Limited Diffusion of ICT Knowledge among Staff

To most SME staff, ICT is just another tool that can be used to enhance their productivity, allowing them to perform their work tasks faster and easier. As most of the staff, however, is under time pressure to perform their day-to-day tasks, they are left with very little time for gaining additional ICT knowledge. Most SME staff have in fact very limited knowledge of IT and, given the time pressures already on them, a very limited or diminishing desire to increase their know-how in that area. Most people are content with what they already know as long as it is enough for them to perform their work tasks.

Considering the limited ICT knowledge and desire of staff to expand it, there are concerns with regards to the uptake of new technologies and rate of adoption of m-transformation. Very few staff will see added value in the new mobile technologies at the beginning. For most, these systems will initially be yet another thing they must learn about, adding onto their existing workload.

Decision Making is Quick and Reactive

The fact that today’s business environment is ever changing is well accepted. Even more so

is the fact that fearsome competition demands fast-paced action and decisions to be handed down as quickly as possible (Ginige, 2004). This is of course all true with any company, but the need to deliver faster, be more agile, and respond to signals from customers as soon as possible is most pressing for SMEs. Those increased pressures lead SMEs into a mode of decision making that is very quick, based on a limited set of facts and appearing very reactive.

Unlike larger organisations, where decision making is a formal and rather rational process, in an SME, most decisions are made by the owner/proprietor who tends to be involved hands-on with the company operations.

In our experience, decision making in the context of SMEs is influenced by the personality and character traits of individuals. Individual personalities, for instance, may prop in the m-transformation decision-making process.

Each Client is Very Highly Valued

SMEs place a lot of value on their clients and are very sensitive to their wants and needs. Although clients are very important for organisations of all sizes, this is particularly true for SMEs. They lack global reach and generally have to compete with each other to maintain and grow their client base from a limited pool of prospects.

Therefore, the client primarily dictates the business, and their voice is heard. Views of clients on mobile business processes are critical in attempting m-transformation. M-transformation aims at providing the SME with greater efficiency while enhancing their clients’ experience in dealing with them. It is therefore very important for SMEs to closely monitor the needs of their clients and implement the right mix of technologies in their m-transformation to effectively fulfil those needs. They cannot afford to leave their clients outside m-

transformation decisions, as the risk of dissatisfying them is imminent and great.

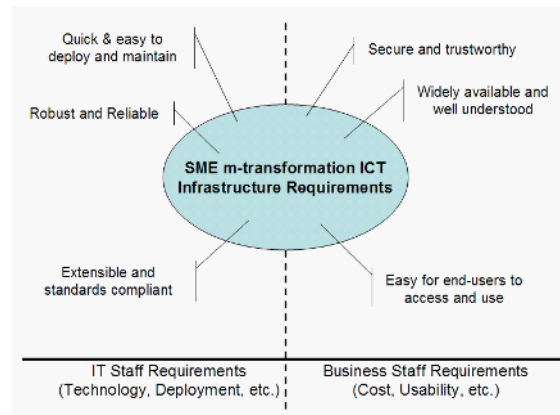
As per the working definition that we have adopted, m-transformation has two sets of prerequisites: those of business process adoption and those of adoption of suitable mobile technologies and products. To coordinate the introduction and mix of these two prerequisites throughout the process of m-transformation, there is also a need for an overarching methodology that can act as a roadmap and guide to the process. The unique characteristics of SMEs that were discussed previously will certainly impact on the business process adoption, the technology mix and adoption, as well as on the methodology itself. The next section discusses how these characteristics of SMEs affect the entire m-transformation process and its constituent parts while describing the requirements those parts have to meet in order to be effective towards the m-transformation process of an SME organisation.

ICT SYSTEMS REQUIREMENTS FOR SME M-TRANSFORMATION

Stemming from our critical analysis and close work with SMEs, we have derived a set of requirements that ICT systems for m-transformation should adhere to in order to be used effectively by SMEs. Meeting these requirements means that they can better fit with the SME way of thinking and approach towards m-transformation, hence being effective to address some of the particular characteristics of SMEs that were discussed previously. These requirements for ICT systems to assist in SME m-transformation are summarized in Figure 2.

As Figure 2 shows, there are six different requirements that ought to be met.

Figure 2. M-transformation IT systems requirements for SMEs



Solutions Quick and Easy to Deploy and Maintain

Solutions for m-transformation suitable for SMEs must be relatively straightforward and quick to deploy while requiring very little maintenance, if any at all. Whilst it is fine for a large company to choose to write custom software or standardise its IT system on particular languages and databases seeking peak performance, these things matter little to SME users. A lot of researchers and many large organisations see it quite fit and natural to build solutions for mobility based on XML (Naedele, 2003), Web services (Staab et al., 2003; Dustdar, Gall, & Schmidt, 2004), and large Java frameworks (Kawashima & Ma, 2004). While this is quite fitting for larger organisations, in the context of an SME, these approaches are mostly unusable due to their complexity and cost.

Secure and Trustworthy

Whatever solutions are adopted to fulfil the technical part of the m-transformation must be secure and trustworthy. Each ICT system an

SME adopts will typically handle a significant part of their business needs, hence security is paramount. This includes security at both levels, internal and external. Every system should be capable of preventing external unauthorized access to itself and its data. It should also however have safeguards in place to limit the access of its current trusted users to the designated data and application areas only. Any attempts a user makes to access information beyond their privileges should be logged. Subsequently, these exception logs should be readily available to designated staff who can take action accordingly. In addition to this, ICT solutions must also have built-in mechanisms of enforcing integrity of their data and prevent unauthorized changes to access logs, security logs, and other such system information. This feature will make them trustworthy and provide a lot more credibility to their output information and reports.

Robust and Reliable

ICT systems to assist in the m-transformation process are bound to be heavily used and relied upon. As such, in order for users to gain confidence in the system which will in turn guarantee its continuing use, it has to be reliable. There is little that can be done to draw users into using a system in the long run if they perceive it as unstable or erratic. Also for users to lose confidence in a system is much easier and faster than for those to gain the same amount of confidence. Beyond appearing reliable, ICT systems must also be robust. By this we mean that they should cater for incorrect use and be prepared to deal with end users that are not sufficiently trained in using the system. Instead of cryptic error messages or crashes, or worse still giving out the wrong information to users, the system should never trust user input, provide informative error messages, and in general

expect the unexpected from its users as much as this is practically possible. A robust and reliable system that covers only some basic functions has far more chances of success in an SME organisation compared to a system that is feature packed with a lot of pleasing user interfaces, but which fails often or demonstrates an erratic and inconsistent behaviour.

Extensibility and Standards Compliance Are Paramount

IT systems developed or purchased must be extensible, able to cope with change, and standards compliant. This is particularly important for SMEs since they do not have the know-how or funds necessary to develop or keep maintaining proprietary solutions. Whatever systems they adopt towards their journey into m-transformation should be based on widely accepted standards, so that the chances of future interoperability with other systems are much higher. Also because of the changing business environment, requirements for mobile access to applications and data are also bound to change often. It is therefore crucial that the systems an SME uses, no matter how simple they are, can be extended to accommodate for those changes over time.

Widely Available and Well Understood

Stemming from the fact that in-house IT expertise is scarce in the SME space and the cost of engaging external experts can be prohibitive for some SMEs as well, it is important that IT systems selected to assist with m-transformation are widely available elsewhere and well understood. It stands to reason that the more commonly used a piece of technology is, the more know-how will be readily available for it in the market, hence at an affordable price. On

the other hand, less known or used technologies getting expert support might prove to be quite a costly exercise. An example of technologies that we feel meet these parameters are the Linux vs. the netBSD operating systems. On one hand, Linux is widely used and gaining in momentum every single day, with a very significant following of IT experts backing it up. This means that finding a consultant that understands Linux and can provide maintenance or other services around it is not very difficult at all, and because of the number of people available, the hourly rates for such consultants are very reasonable. On the other hand, netBSD is another free and open source operating system that in many respects provides superior performance than Linux out of the box. It has a significant number of followers as well which, however, are much less compared to those for Linux. For an SME to get by netBSD expertise would be a hard task, especially when they are in a hurry or in the event of a system problem. Accordingly, even if they find an expert in netBSD, it is expected that the cost for such a person would be higher than the equivalent one for a Linux system.

ICT Solutions that are Easy to Use

Whatever solutions are adopted to fulfil the technical part of the m-transformation have to be easy to use. The more common the technologies used are, the easier it is convincing an SME to embrace them since the perceived risk stemming from their usage is reduced and their effectiveness largely proven. Also the training burden will be less, and for very common technologies and systems, it is possible that staff might already be well versed in their use from past roles they have had. Such an example is the Microsoft Office suite of software, for instance, where most sales and administrative staff would typically have enough knowledge

of how to use it to be productive with it straight away. For less common systems, we found that integrating additional functions onto existing applications that staff are using on a daily basis makes for a very quick way to have this functionality adopted by end users. For instance, in most of the SMEs we work with, a simple e-mail to the SMS gateway system would boost productivity significantly, allowing for faster, reasonably priced communication that most staff could use with virtually no training at all. On the other hand, offering SME staff a state-of-the-art PDA to use over GPRS to access their company's systems might prove an error-prone, complex undertaking with a lot of associated costs which could also face resistance from end users since they have to get significant training in it before they can be productive using it.

Technology aside, business process adoption is the second prerequisite ingredient in a successful m-transformation, and it is equally affected by the unique SME characteristics identified previously. The section that follows discusses the requirements for business process adoption during m-transformation.

BUSINESS PROCESS ADOPTION REQUIREMENTS FOR SME M-TRANSFORMATION

M-transformation is typically a long process during which an organisation undergoes quite a few large and several smaller changes. These changes include, of course, ICT systems and infrastructure, but also they relate to business process changes and the way staff go about performing their daily tasks.

As such, it is important to establish the requirements to be met for the smooth and successful adoption of business processes by SME organisations undergoing m-transformation. These requirements are discussed below.

Small and Incremental Changes

It is generally true that the big-bang approach works for only very few things when it comes to dealing with SME organisations. Because of their size and capabilities, SMEs tend to be very much influenced by change. Hence, adopting new business processes that allow for mobility has to be done as a series of small and incremental changes, as opposed to the big-bang approach. Business processes in SMEs tend to be long standing, and they change only when they absolutely have to. In our experience working extensively with SMEs, it is very rare that they will pursue best practice or that they will knowingly change business processes of theirs that “work” for better, more efficient ones. For most companies their current processes are the result of years of evolution and trial and error; expecting therefore that they will abandon those overnight for mobile-enabled processes is fairly unrealistic. When on the other hand changes are introduced slowly and are small enough to be easily understood by staff, while at the same time compliment the existing process and not upturn it, they are far more likely to succeed.

Adoption Must Minimize Risk by Changing Non-Core Processes First

Another requirement for adopting business process changes during m-transformation is that these are done to non-core business processes first. As the case is with every type of change, the exact results of the change are unknown until it is done. In some cases the effect can be predicted, but when it comes to business processes, it is best that the risk is minimised. This can be done through the introduction of mobile capabilities to non-core processes first. Support processes, as they are called, can provide

a great test bed for putting the technology through its paces, as well as allowing staff to get accustomed to change in the way things are done, while realizing the benefits of mobile technologies. As staff become accustomed to the new technologies and the abilities of performing tasks faster and smarter with increased flexibility, they are more likely to welcome changes to some of the more critical, core business processes. Also by the time those get to be adopted to a mobile-enabled environment, the return on investment for m-transformation typically has already started materialising as well, proving a further incentive for continuing with m-transformation.

Change Introduction Must Be Gradual and Reversible

Another important requirement for the process of adopting new business processes during m-transformation is the ability to be gradual and the changes reversible. SMEs tend to place a lot of value on their existing processes and be attached to them. Therefore, gradually introducing the changes allows staff to settle into the new routine of doing things, as well as over time detach themselves from the old habits. Additionally, as yet another means of reducing risk, both real and perceived, changes should be reversible. This means that business process changes ought to be well defined and measurable, so that if in the course of monitoring the change it turns out that it is underperforming or has bad side effects, it can be reversed, leaving things to their previous working state. This is nearly commonplace in larger organisations; however, in the SME space it is very often overlooked. It is frequent in our experience for the management of the SME to take drastic, nearly ad-hoc decisions under pressure in response to an adopted business process that is under performing. In those scenarios, instead

of reversing the process to its predecessor, they often try to “patch” it, with even worse results sometimes.

Return on Investment Must be Continuous and Apparent

The discussion of the SME characteristics made clear that they like a large, dedicated budget to put towards their efforts of m-transforming. Hence, every dollar spent towards that goal has to offer and also be seen to offer value as soon as possible. To that end, adopting existing business processes to work in a mobility-enabled environment is no different. It is essential therefore that new or improved business processes offer continuous and apparent return on investment (ROI). It is hard to keep the SME management interesting in perusing the m-transformation goals if they see no obvious value stemming from the changes their business processes are undergoing along the way. This is particularly true with the adoption of business processes, given that they are intangible—unlike PCs or networking gear, for instance, which people can readily see and touch as soon as they are acquired and deployed.

Change Should Gradually be Applied throughout the Organisation

While trying to minimise risk, we identified the requirements of business process adoption for mobility to target non-core processes first. Another requirement now is added, that of ensuring that changes to business processes will cascade over time to reach as many staff members across the organisations as possible. The reason for this requirement is to ensure that there is enough penetration and visibility of the changes that are happening towards m-transformation throughout the staff population.

Without this taken into account, business process adoption and therefore m-transformation may be perceived as something smaller than it really is, affecting only a handful of staff—especially when it is first applied to secondary business processes that are typically inside a single department or between a handful of individuals within the company. On the contrary, the changes have to eventually cascade down and flow onto as much of the staff as possible. This way the benefits of m-transformation are more readily realised, while at the same time everyone gets to see the true impact that m-transformation is making across the organisation, hence gaining more buy-in and support at all levels.

METHODOLOGY REQUIREMENTS FOR M-TRANSFORMING SMES

So far we have discussed the requirements for ICT systems and the adoption and change to business processes as part of m-transformation as they are dictated by the SME characteristics. There is however one more component that must be looked at very carefully when dealing with SME m-transformation. This is the methodology. The methodology for SME m-transformation is very important because it is what will “glue” together the technology on the one hand and the new or improved business processes on the other to produce the desired set of outcomes in the end—namely, increase the efficiency of the organisation, offer better management and utilization of resources, and allow for more responsive customer service. As the case has been with both technology and business process adoption, the SME characteristics also impact the methodology itself. Therefore, it needs to also meet particular sets of requirements in order for it to be well suited to the task within the SME space. Some of the

most important requirements are shown as follows.

Staged, Well-Communicated Process

Taking an SME from its current state to being fully m-transformed will certainly take time. The process duration can vary depending on the size and awareness of the SME, its level of technical competence, and many other factors. What is certain however is that m-transformation is not going to happen overnight for anybody. To that end, the methodology must adopt an overall staged process that leads to full m-transformation. For instance, there is no need to try and implement the entire infrastructure that has to be put in place all at once. In fact doing so is a sure way to fail quickly. Instead a gradual ease into the new technologies would be the best path to their effective adoption and use. The same holds true of course for the business process adoption as we have already discussed. The methodology must also allow for communicating the process in advance with the stakeholders in the organisation. M-transformation requires quite a few resources and continuous commitment to it in order to succeed, hence people ought to be kept informed about what is going to happen next before they are expected to embrace change and promote it.

Benefits of M-Transformation Must be Shown Quickly and Continuously

Very much like the case is for ICT systems and business process adoption, the methodology for m-transformation as a whole should ensure that the benefits of the efforts can be easily seen and that they start to come into existence as early into the process as possible. If results are not shown soon, people are bound to lose faith

and confidence in the process, as well as lose sight of the end goals of m-transformation that will only materialise after a relatively long time and through persistent investment.

Methodology Must be Balanced Between ICT and Business Processes Adoption

In order for the methodology to be suitable, it has to offer a balanced approach between technology adoption and business process re-engineering or adoption. The two prerequisites of m-transformation must be kept in sync, therefore the methodology has to have provision for monitoring these two very important items and track their progress over time. The actual ratio or “mix” between the two will certainly always be different, as it depends on a lot of parameters like the current levels of technology adoption within the SME and their natural capability to respond to changes well or not. By applying a methodology that does not make provision for measuring and tracking the amount of each ingredient used, it is nearly impossible to keep track of the current mix, let alone optimise it to better suit the particular SME undergoing m-transformation.

Plan for the Worse and Expect Resistance to Change

Humans have a natural tendency to resist change and in many situations will go out of their way to ensure that the status quo is maintained. This is also something that might happen during the journey of m-transformation inside an SME. Some people will be receptive to change, while others will be less so and typically a minority will actively oppose any kind of change. The reasons behind the resistance to change are many and diverse, and we shall not cover them here. Suffice to say that the m-transformation

methodology has to be built taking the worse case scenario in mind, as well as the fact that people shall resist changes necessary for the full implementation of m-transformation. Therefore the methodology has to be flexible to deal with such situations and provide multiple levels of fallback in order to effectively manage those accordingly.

Methodology Evolution Must be Based on Feedback Received

Just like m-transformation is a dynamic process that over time has to adjust to the changes of the environment until it helps the organisation reach its longer-term goals, so should the methodology for m-transformation be as well. More specifically, the methodology must have built-in feedback loops where input from the SME itself should be actively solicited and used to adjust the future steps or timing of those as prescribed by the methodology for m-transformation. If the methodology was set in stone and was completely isolated from the environment, it would have very few chances of actually being successful. The requirement for gathering, analysing, and incorporating feedback into the next steps of the methodology along the entire m-transformation journey guarantees that

changes in the environment will be taken into account, hence increasing the chances of success for the entire process of m-transformation.

The section that follows describes the actual steps in the methodology for m-transforming SME organisations.

OUTLINING THE METHODOLOGY FOR M-TRANSFORMING SME ORGANISATIONS

After discussing the requirements of the methodology for SME m-transformation, we will now move on to outlining the actual steps involved in the methodology itself. As Figure 3 shows, there is a total of eight steps involved and a prescribed sequence in which those steps need to be taken. You will also note that some of these steps are repeated more than once throughout an m-transformation project.

With Figure 3 in mind, setting the overall framework for when and where each of these steps is taken, let us now look into each of the steps in more details.

Identify the stakeholders. The first step in the process is to identify the stakeholders in the SME organisation. In other words, identify and establish good working relationships with the

Figure 3. Considerations in selecting ICT for SME m-transformation

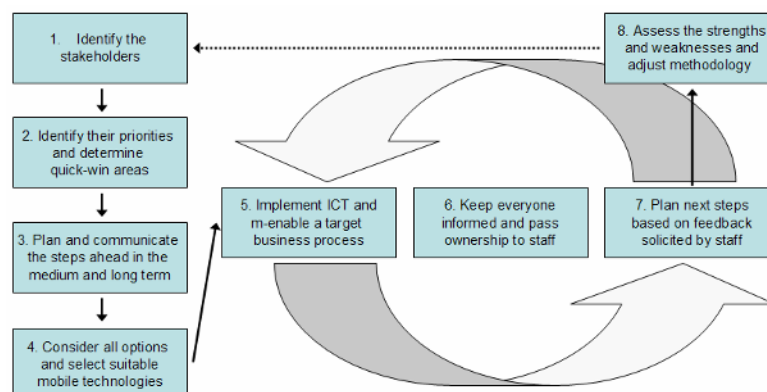
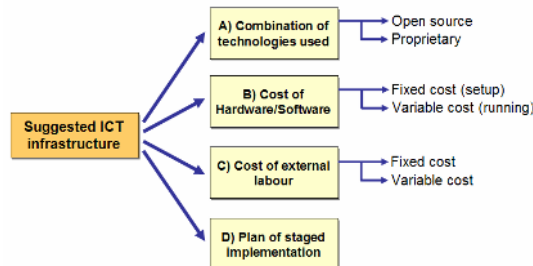


Figure 4. Considerations in selecting ICT for SME m-transformation



key staff in the company that can provide you with information and support for the m-transformation process. Some of these people of course would include the managing director and owner(s) of the company. In addition to those, other key staff might include the financial controller or chief accountant of the company or perhaps the production manager if the SME is involved in the manufacturing sector.

Identify their priorities and determine the areas where mobility could offer a quick win or significant competitive advantage. Via interviews with the key stakeholders identified previously, one should be able to identify a few important pieces of information, including the priorities these people have (and collectively the organisation has) in regards to areas in which they believe the introduction of mobile-enabled technologies will be more appropriate. Also through understanding more about the organisation, one could determine what particular processes, departments, or other areas of the organisations could realistically offer a good initial test bed for trialling the changes and technology associated with the m-transformation effort. These areas or processes would ideally be able to provide either a quick-win situation of a significant competitive advantage once they become mobile enabled.

Put together a medium- to long-term plan of the steps involved in m-transforming this

particular SME and make it available to all staff. This is an extremely important step that should be noted accordingly. It puts the foundation of establishing channels of open communication and information dissemination throughout the SME, allowing as many people as possible to be informed and also get excited about the upcoming e-transformation. Ideally these steps should establish m-transformation as the process through which the company will realise some major benefits as opposed to a black box process that involves a lot of time and money investment whilst offering unclear or undefined results.

Consider all options when selecting suitable mobile technologies. The selection of technologies capable of fulfilling the ICT requirements for m-transformation is an important step. During this, one should consider all options within the SME budget and decide which combination of technologies will provide the best value, while meeting the requirements set out previously with regards to ICT systems. In addition to those requirements, Figure 4 also offers a comprehensive overview of additional things that would need to be considered before arriving at a final decision for the technology mix of choice.

Implement the technology infrastructure selected in an incremental manner while using these added facilities to enhance aspects of the business processes previously identified as quick-win situations. This is one of the key steps that will repeat over and over for each targeted business process during the entire course of m-transformation for the organisation. In this step the theory and planning done so far is put to the test. The ICT framework is established or enhanced, and the added functionality is used to improve existing business processes or replace old ones with new, more efficient processes. It is also during this step that resistance will be felt from some

of the staff, since now they would realise that m-transformation is actually affecting how they do or they used to do things, as opposed to being just an initiative on paper with no actual implications to their roles and duties. Because this step will be repeated for as many business processes as will become mobile enabled, ongoing management buy-in is crucial along with wide acceptance of the changes and progress by the majority of staff.

Ensure everyone is kept informed and takes ownership of the process as appropriate. Key to the ongoing success of the effort towards m-transformation is keeping everyone informed, making progress visible, and also passing the ownership of certain subtasks involved in m-transformation to staff of the organisation directly so that they can relate to those and push them further ahead. Although we already mentioned this numerous times before, we cannot stress enough the importance of establishing contact and a communications channel to the staff of the SME organisation through which they can find out about the progress of the m-transformation process as a whole. Without this, people will quickly lose sight and faith in m-transformation which, in turn, will have a detrimental effect to it.

Solicit feedback and use it in planning the next steps towards the goal of m-transformation. This step is also one that repeats over and over in the methodology, and it is important to do so in order to ensure that the next few steps planned will be effective and carry the entire process forward. There are numerous ways of getting the feedback; what is more important to understand at this stage, however, is that there is a need to collect feedback from staff at all levels and incorporate it into the future planning of activities and tasks in the m-transformation journey.

Assess the strengths and weaknesses of the project. At the completion of the process, carry out an assessment of what the strengths

and weaknesses of it were so that they can be provided as feedback in the overall model of SME m-transformation and improve the methodology over time, as well making it even more suitable for the task at hand when engaging with a new SME wishing to undergo m-transformation. This step will help ensure that the methodology steps are kept in sync with the actual needs and set of parameters that affect SMEs in their journeys to m-transformation. That way the methodology can better serve the cause and also grow and evolve over time, increasing in its completeness that will hopefully lead to it leading to better results for subsequent m-transformation projects.

Armed with the actual methodology steps, the next section will help you put them in context by offering some insights we gained from applying this particular methodology to a few different SMEs that we are closely working with.

PUTTING THE METHODOLOGY INTO ACTION—EXAMPLES FOR SUCCESSFULLY M-TRANSFORMING SMES

Through the years we have worked with many SME organisations, first assisting them with their e-transformation journey and later with their m-transformation. Through these engagements we have learnt quite a few valuable lessons. In this section we give some brief insights to the things we have learned in the hope that you will not have to learn to avoid those the hard way, but you will rather know of some of the pitfalls ahead of time. Here are some of the lessons learned, along with a brief outline of each one.

Do not rely on management buy-in alone to promote m-transformation. Even though you have heard us talk about management buy-in for m-transformation in numerous places

throughout this chapter, one of the things that we found out in practice was that on its own it is not enough. Management buy-in is essential in every way for m-transformation to start and keep going. If however there is no buy-in of e-transformation from the wider staff base, it can be seen as yet another management initiative “pushed down peoples’ throats” and lead to increased resistance against it from staff members. This is why the methodology puts so much emphasis on gaining the management buy-in up front, but also open and maintain an open channel of communication with all staff, not only the senior management of the organisation.

Give ownership of the e-transformation process to people, and help them progress it further. We have already spoken of the need to pass the ownership of individual business process adoption on to staff members, and allow them to provide feedback and contribute to their future shape. What we also found to be very important is the need to, as a whole, push the ownership of e-transformation on to staff within the company. In cases where we acted just as a mediator between technology and business users, with the latter carrying the ownership of m-transformation, the results have been very good. On other hand, acting as a gatekeeper and protecting the overall process while pushing the ownership of individual processes only to people can have the wrong effect. It sometimes sends the message to staff that they are not trusted to handle the larger process of m-transformation and that they are instead micro-managed throughout the entire journey.

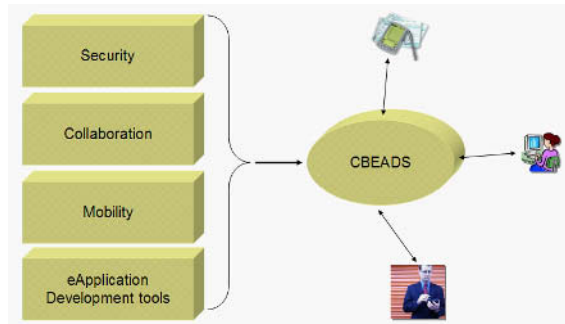
Keep a balanced position against technology and business users throughout. Depending on your background, whether technical or more business oriented, you may have a natural tendency to favour either the technology or the business process end of things in m-transformation. While this is understandable,

you should refrain from expressing this while engaged in an actual m-transformation. If IT or business staff perceive you as being partial to the other one respectively, your credibility might be damaged in their eyes. This will of course lead to making the remaining m-transformation journey more difficult. Therefore, even though the mix of technology vs. business process adoption is not always equally divided, your attitude towards approaching each should be that of a person well informed, but open minded, without preconceived ideas about either aspect.

Select technologies based on realistic value, not current trends or hype. Selecting the ICT framework necessary to power the efforts for m-transformation—and in particular for an SME with limited knowledge of IT and very small budget for such purchases—can be a real challenge. From our experience it is best to look at solutions that are within the SMEs’ budget and offer the best value for the money. Even though the IT industry puts a lot of emphasis on current trends and “hot” technologies from time to time, you should not let hype drive your decisions. Remember that these decisions can make or break the chances of success of m-transformation for the particular organisation.

Look for components-based technologies that can expand and scale. Selecting technologies suitable for m-transformation can be a very hard task in itself. The sheer breadth and depth of solutions currently available both as commercial software and at the state of research ideas and prototypes is very substantial. In jointly working with SMEs, we looked at several available solutions—commercial and others—ranging from simple gateways between e-mail and SMS or MMS over the common mobile phone networks, all the way to new environments for mobile collaboration over specific domains (Dustdar & Gall, 2003; Litiu &

Figure 5. CBEADS© high-level architecture



Zeitoun, 2004; Looney & Valacich, 2004) and mobile disconnected business processes (Sairamesh, Goh, Stanoi, Li, & Padmanabhan, 2002). One thing that we found missing was a common framework that could bring together the disparate set of technologies that are needed in order to provide enough IT facilities to better business processes and in turn reach m-transformation.

As an answer to this gap, and to fill it while assisting several SME organisations in achieving their m-transformation goals, we have developed and successfully used a home-grown technology framework known as CBEADS©. The name stands for Component-Based E-Application Development (and deployment) Shell. As its name suggests, CBEADS© is component based and can be extended easily to accommodate changes in the requirements from one engagement to the next. Being fully Web enabled allows for quick and easy deployment, while all the user interfaces are put together via HTML controls that end-users are familiar with through their general use of the World Wide Web. CBEADS© is multi-platform and so far is supported on both MS Windows™ and Linux operating systems.

Through CBEADS©, the SMEs we have worked with covered their needs for a central framework for deploying their e-business mobile-enabled applications, as well as an environment to develop and extend those. With user management and centralized authentication and access control services also offered by CBEADS, the overheads associated with creating and deploying new applications to select groups of staff in the organisation are largely alleviated.

Figure 5 shows the four major subsystems that make up the CBEADS© framework and how it provides a consistent, device-independent view of the applications and data it has access over.

When selecting technologies for m-transformation, it is very important to ensure that they meet some common characteristics. As is the case with CBEADS©, you should also strive to establish a common platform for managing users and access control to system applications. Managing users and groups in one place greatly reduces overhead and makes the process of deploying applications significantly faster. Also a World Wide Web-accessible system should be preferred over a client-server one, again due to deployment being easier and not dependant on the operating system of the client machines, which in SMEs could vary significantly. Finally, having an integrated development and deployment environment can prove helpful when small changes are needed to existing applications or extensions are required for applications to become mobile enabled.

We hope that these additional guidelines stemming from our experiences in working with SMEs undergoing m-transformation will prove useful and assist others as well to carry out the process with great success.

CONCLUSION AND FUTURE DIRECTIONS

The next logical step from an organisation's e-transformation is that of m-transformation, which proves to be particularly difficult to succeed in especially in the case of SME organisations. SMEs have unique characteristics that set them apart from larger companies. These characteristics impact upon all aspects of m-transformation, including the requirements for a suitable mix of ICT solutions, the adoption of business processes that can take advantage of the new technologies, and the methodology of m-transformation as a whole.

We have presented the individual requirements for each of these aspects of m-transformation along with reasons for each to succeed in the SME space. Stemming from these we looked at the exact steps involved in the methodology for m-transformation applicable to SME organisations in particular. Finally, from our experience in helping SMEs m-transform, we have offered a set of guidelines on how to select technologies that could assist in the process of m-transforming an SME, as well as how to deal with management buy-in and more.

In view of all the benefits—both tangible and intangible—that m-transformation promises to deliver to its embracers, it should come as no surprise that it is an already “hot” topic that will no doubt continue to gain in popularity for quite some time. Hopefully, this chapter has armed you with the necessary understanding of the unique terrain of the SME landscape in relation to m-transformation. The list of requirements we presented may prove useful to SMEs themselves, software vendors, and systems integrators who are involved with assisting companies along their m-transformation journey.

REFERENCES

- Arunatileka, S., & Ginige, A. (2003, June 3-6). The seven e's in e-transformation—A strategic e-transformation model. In *Proceedings of the IADIS International Conference*, Lisbon, Portugal.
- Dustdar, S., & Gall, H. (2003, February 5-7). Architectural concerns in distributed and mobile collaborative systems. In *Proceedings of the 11th Euromicro Conference on Parallel, Distributed and Network-Based Processing*, Genova, Italy (pp. 475-483).
- Dustdar, S., Gall, H., & Schmidt, R. (2004). Web services for groupware in distributed and mobile collaboration. In *Proceedings of the 12th Euromicro Conference on Parallel, Distributed and Network-Based Processing*, A Caruna, Spain (pp. 241).
- Ginige, A. (2002). New paradigm for developing evolutionary software to support e-business. In S. K. Chang (Ed.), *Handbook of software engineering and knowledge engineering* (Vol. 2, pp. 711-725). Hackensack, NJ: World Scientific.
- Ginige, A. (2004, March 26-27). Collaborating to win—Creating an effective virtual organisation. In *Proceedings of the International Workshop on Business and Information*, Taipei, Taiwan.
- Ginige, A., Murugesan, S., Khandelwal, V., Pollard, T., Costadopoulos, N., & Kazanis, P. (2001). *Information technology in Western Sydney: Status and potential*. Western Sydney: University of Western Sydney.
- Kawashima, T., & Ma, J. (2004, March 23-24). Tomscop—A synchronous p2p collaboration platform over JXTA. In *Proceedings of the*

24th International Conference on Distributed Computing Systems, Tokyo, Japan (pp. 85-90).

Kazanis, P. (2003). *Methodologies and tools for e-transforming small to medium size enterprises*. Unpublished PhD, University of Western Sydney, Australia.

Litiu, R., & Zeitoun, A. (2004, January 5-8). Infrastructure support for mobile collaboration. *Proceedings of the 37th Annual Hawaii International Conference on System Sciences*, Big Island, HI.

Looney, C. A., & Valacich, J. S. (2004, January 5-8). Mobile technologies and collaboration. *Proceedings of the 37th Annual Hawaii International Conference on System Sciences*.

Marmaridis, I., Ginige, J. A., & Ginige, A. (2004a). Web-based architecture for dynamic e-collaborative work. In *Proceedings of the International Conference on Software Engineering and Knowledge Engineering*, Banff, Canada (p. 445)..

Marmaridis, I., Ginige, J. A., Ginige, A., & Arunatilaka, S. (2004b, May 23-26). Architecture for evolving and maintainable Web information systems. In *Proceedings of IRMA04*, New Orleans, LA.

Marmaridis, I., & Unhelkar, B. (2005). Challenges in mobile transformations: A requirements modelling perspective for small and me-

dium enterprises. In *Proceedings of the 4th International Conference on Mobile Business (ICMB 2005)*, Sydney, Australia (pp. 16-22).

Naedele, M. (2003). Standards for XML and Web Services security. *Computer*, 36(4), 96-98.

Ranjbar, M., & Unhelkar, B. (2003, December 16-18). Globalisation and its impact on telecommuting: An Australian perspective. In *Proceedings of the International Business Information Management Conference (IBIM03)*, Cairo, Egypt. Retrieved from www.ibima.org

Sairamesh, J., Goh, S., Stanoi, I., Li, C. S., & Padmanabhan, S. (2002, September 28). Self-managing, disconnected processes and mechanisms for mobile e-business. *Proceedings of the 2nd International Workshop on Mobile Commerce*, Atlanta, GA (pp. 82-89).

Staab, S., van der Aalst, W., Benjamins, V. R., Sheth, A., Miller, J. A., Bussler, C., et al. (2003). Web Services: Been there, done that? *Intelligent Systems, IEEE [see also IEEE Expert]*, 18(1), 72-85.

Unhelkar, B. (2004, December 15-18). Globalization with mobility. In *Proceedings of the 12th International Conference on Advanced Computing and Communications (ADCOM 2004)*, Ahmedabad, India.

Chapter XLI

Business Process Mobility

Harpreet Alag
Agilisys Limited, UK

ABSTRACT

This chapter introduces the concept of business process mobility. Mobility in this case refers to the ability of a human resource to work from multiple locations and in non-office environments; business process mobility involves enabling that resource to carry out specific aspects of a business process while mobile. It attempts to explain where and how mobile enabling processes and systems can benefit. The chapter argues the need for redesigning business processes to support mobility instead of simply adding mobile systems. It further attempts to explore the approach for analyzing and redesigning processes to support mobility. The author also hopes to provide an understanding of mobile systems and their role in enterprise mobility. The chapter touches upon the essentials of mobility strategy and concludes by discussing key contents for a business case for mobile enabling business processes.

INTRODUCTION

In the last decade, advancements in technology, particularly the Internet and electronic commerce, have changed the way people work and live. Technology has helped businesses around the globe to bring about drastic improvements in the how they conduct business, and helped them produce new products and services faster than ever. Some of these products

and services have changed the face of industry completely, for instance Internet banking.

Mobile technologies have been at the forefront of all technological developments in recent years. The growing number and use of mobile phones and personal digital assistants (PDAs) is a good indicator of overall potential of mobile applications. Globally, the number of mobile phone users reached 1.5 billion in June 2004 (IT Facts, 2005). Mobile phone sales in

2004 increased by 30% to 674 million units compared to the sales in 2003, while the sales growth in 2003 compared to 2002 was only 20.5%. The increasing sales of mobile phones also include replacements or upgrades. A wide range of mobile devices and applications are available in the market, including laptops, handheld computers, tablet PCs, PDAs, and mobile phones with PDA capabilities. Some of the leading names in PDAs are PalmOne, HP, RIM, and Dell. A number of devices having wireless capability is rapidly increasing. In 2004, 44% of PDA devices offered integrated wireless network support (IT Facts, 2005). Some of the popular mobile applications are personal information management applications such as contacts, calendars, and more recently wireless e-mail. Short Messaging Service, popularly known as SMS or “texting”, is still the most widely used mobile application. Convergence of Internet and wireless technologies—also known as “wireless Internet”—is considered the fastest growth area in technology industry. The possibilities offered by this convergence are virtually unlimited. Large-scale use of mobile technologies is expected to have a large impact on business and consumers in the coming years.

This chapter focuses on the use of mobile applications in core business processes across an enterprise. It is an attempt to understand how mobile applications can be used to provide mobility to business processes and extend the philosophy of mobility to build a mobile enterprise.

The following topics will be discussed in detail:

- Evolution of mobile business applications
- The concept of mobile business process
- Motivation for mobile enabling business processes
- Enterprise mobility and enterprise mobile systems

- Redesigning for mobility
- Enterprise mobility strategy
- Business case for mobile enabling

EVOLUTION OF MOBILE BUSINESS APPLICATIONS

The first generation of mobile applications worked in a disconnected or off-line mode: they were synchronised with a desktop computer by physically connecting using a USB or serial port connection. Applications available on compact mobile devices ranged from personal information management (PIM), basic word processing, and spreadsheets, while laptop computers offered almost the same capabilities as desktop computers, but without mobile connectivity. The first generation of applications helped to some extent in improving individual productivity by providing users with the ability to do basic tasks even when away from the workplace. This only worked well with personal information such as contacts and calendars, as data integrity was not at risk.

With the next level of evolution in mobile communications, it became possible to build and implement mobile applications that operate in near real time. The second generation of applications worked in an online, but not always-on mode, and could transfer data in near real time. These applications were basic in functionality due to limitations of mobile connectivity, hardware capability, and data transfers, and relied primarily on wireless application protocol (WAP) and Short Messaging Service. Such applications included checking account balances via mobile phones, receiving alerts/notifications, and checking order status. These applications mostly obtained data from the Internet or a specific server and presented it in a suitable format on mobile devices by means of device-specific user interfaces. Such

applications mainly worked by requesting and retrieving information from enterprise systems and relaying it on mobile devices using specific communications standards/protocols. The use of mobile applications was largely restricted to non-core activities for productivity gains and not core business processes. Calendar, contact, and time management are some of the popular mobile-enabled non-core activities. The reason for this is that core processes in a business can be very complex in terms of number of activities, decision making, and number of people involved. Introducing a mobile element in the IT systems of such processes was often not feasible due to one or more of the following reasons:

- Limited capabilities of technology
- High cost of mobile enabling to address complex requirements of core processes, implying ROI cannot be achieved
- High risk or disruption in implementation

With further developments in mobile infrastructure and proliferation of mobile technology types and providers, the third generation of “always-on” mobile applications is now possible. These applications are always connected in near real time to a network to send and receive data. Even complex business tasks requiring “always-on” connectivity and data transfer can be supported by these applications. Mobile enabling complex core processes of a business have now become feasible and economically practical with “always-on” connectivity and wireless Internet. There are multiple connectivity options that when used in combination can provide an effective mobile solution for most requirements. These include General Packet Radio Service (GPRS), Code-Division Multiple Access (CDMA), WAP, Bluetooth, and Wireless LAN (Wi-Fi). GPRS, a part of the GSM (Global System for Mobile

Communications) family, is fast evolving as the leading standard for wireless data communications around the world. According to Forrester Research, in March 2005, 78% of the total 1.5 billion mobile telecom users worldwide were connected using a GSM network (IT Facts, 2005). Wireless e-mail is by far the most popular mobile application across the world. RIM’s Blackberry™ is one of the leading wireless e-mail applications; it works on automatic push technology. According to Radicati Group, around 3.2 million business users were expected to have used wireless e-mails in 2004. Gartner reported 30 million users were expected to use Wi-Fi hotspot worldwide by the end of 2004, up from 9.3 millions users in 2003 (IT Facts, 2005).

These new-generation mobile applications are increasingly finding their place in core business activities beyond improving personal productivity. Across industries, many business functions are exploring mobile applications to increase business efficiency. Customer relationship management (CRM), supply chain management (SCM), and (physical) asset management are some of the prime business areas that can benefit from mobile applications. In addition to core horizontal processes, mobile applications are also finding success in industry-specific vertical processes such as pharmaceutical sales and insurance. Globally, businesses are exploring mobile technologies to support their mobile workers and strengthen the concept of mobile offices. Growing capabilities and robustness of technology are expanding the potential role of mobile applications in core business. With the growing use of mobile applications, the term ‘m-commerce’ has already found cognizance. Mobile commerce is the process of channelling electronic commercial transactions through mobile devices.

Some common applications in use are listed below:

- “Office” workers checking e-mails and viewing documents on Blackberry™
- Handheld computers with bar code scanning software used to check deliveries in and out of warehouses
- Maintenance engineers downloading their day’s job list and uploading the job completion status

MOBILE BUSINESS PROCESS

The early use of mobile applications in automating business processes was primarily technology driven. In order to realise potential benefits of mobile technology, the design of mobile applications was driven by possibilities offered by technology, namely availability of mobile devices and mobile connectivity. Process managers changed their processes to utilise technology options available, rather than ensuring that technology worked within their business requirements and achieved all the business objectives. However, with the latest technological developments, processes can be mobile enabled using a business-driven approach, where mobile technology is just an enabling tool.

The term “business process” has been defined by a number of authors including popular definitions by Michael Hammer and Thomas Davenport. Davenport’s definition is perhaps simplest, according to which a business process is “a specific ordering of work activities across time and place, with a beginning, an end, and clearly identified inputs and outputs” (Davenport, 1993). Business processes can be complex and large scale or smaller processes made up of few activities. A good example of a high-level process is procure-to-pay, which covers

end-to-end process, beginning with requisitioning and ending with the payment for goods received. Completing an expense form is a common example of a low-level activity applicable to all businesses. Processes can be decomposed to smaller sub-processes and viewed at lower levels of detail. The lowest level sub-process is an activity with a well-defined input and output. Gruhn and Kohler (2003) proposed the term *mobile business process*, according to which mobility is given for a business process, when at least for one of the process activities there is externally determined *uncertainty of location* (Valiente & van der Heijden, 2002) and the process needs cooperation with external resources for its execution. I will discuss this in more detail in one of the later sections of this chapter. To simplify the above definition: a mobile business process is one that consists of one or more activities being performed at an uncertain location and requiring the worker to be mobile. Such a process can be supported by mobile systems to increase process efficiency. For processes that are supported by mobile systems, the term *mobile-enabled business process* is more appropriate, to differentiate from a *mobile business process*.

MOTIVATION FOR MOBILE ENABLING PROCESSES

This section explores the motivations behind mobile enabling business processes. The key motivation for mobile enabling business processes is the need to serve customers faster and reduce costs. Globalisation and intense competition demand that businesses respond faster to changing market and customer needs, by reducing time-to-market for products and services and serving their customers in a near instantaneous fashion. Businesses are continu-

ously striving to become more efficient and effective. The ever-increasing pressure is only forcing businesses to be more innovative. Businesses considering mobility of processes will have to build a business case for having their processes mobile enabled, justifying how it will help their business. The essence of a mobile-enabled process is that a worker should be able to conduct essential business regardless of their location. The extent to which an enterprise addresses mobility of processes depends on the following factors:

- Need for *mobile enabling processes*, in other words the need to support mobile information systems
- Technological feasibility of the mobile information system
- Financial justification based on cost-benefit analysis model
- Operational feasibility—human factors and so forth

As far as the motivation for mobility is concerned, it can be argued that there are two basic drivers for addressing process mobility.

1. **Process Efficiency:** A set of factors that demand cost reduction, improved customer service, and response time. An example of mobile enabling a process for efficiency is sales staff being able to create online quotations and orders at the customer site using their mobile devices.
2. **Increased Personal Productivity of Employees:** Time and travel management are some of the common processes for mobile enabling and achieving significant productivity improvements.

In some areas of business, the benefits of enterprise systems cannot be fully realised, as a large number of mobile workers cannot ac-

cess wire-bound systems. As a result, many organisations cannot realise the expected return on investments in expensive enterprise systems. In addition to that, most business processes and supporting systems are designed around office-based employees and are not friendly to mobile workers. A good example is an expense claim form, which is typically made available on company intranets and cannot be accessed by sales staff when they are away from the office. Mobility of processes supported by systems can help achieve additional return on investments in enterprise systems such as ERP, CRM, and SCM. Business managers do not introduce new technologies into the business processes to become technology leaders; they are interested in the added business value and how technology contributes to it; in how they can achieve the goals of streamlining their processes. From the point of view of mobility, managers are interested in what value mobile enabling can add to the business processes.

MOBILE ENTERPRISE: PROCESSES BEYOND THE WORKPLACE

Businesses that aim to support mobile workers and enhance process effectiveness will need to consider extending their process and systems beyond the workplace. In order to achieve this, they will have to change their processes and systems in line with the objectives of process mobility. An enterprise that can transform its processes to make its mobile workers and processes more effective can be considered a *mobile enterprise*.

Smart organisations will aim to leverage process mobility for strategic advantages. Such businesses derive tactical and strategic value from mobile enabling processes. In order to

gain maximum benefits from mobility, organisations should have a mobility strategy defined—aligned to its business strategy—and it should complement the IT strategy.

MOBILE ENTERPRISE SYTEMS: SYSTEMS FOR MOBILE PROCESSES

Today, the use of mobile systems is ubiquitous in our daily lives, for instance the courier delivering a parcel and recording delivery details on a handheld device, and the recipient signing on the handheld device using a stylus. The receipt details are instantly sent to the central systems, thus providing real-time status of deliveries for their billing and customer service departments. Restaurants are also benefiting from mobile systems by providing handheld computers on which serving staff input orders and view the status of their orders. Kitchen staff can see the orders instantly on the kitchen computer and update this status when an order is ready. This saves time moving between the dining area and the kitchen, and thus reduces the overall time taken to serve food to the customers. Restaurants and bars using Chip ‘N’ Pin technology for credit card payments have chosen to use mobile devices to save customers needing to walk to the counter to key in their PIN. This is an example of a mobile application requirement arising out of regulatory changes.

Mobile systems can be beneficial across a number of processes in most business areas. Large corporations are embracing mobile systems in almost all major areas of business such as sales, procurement, warehouse management, and so on. A number of mobile solutions available in the market are truly enterprise encompassing in nature as they fully integrate with the existing enterprise systems and bridge across major processes of the enterprise. Com-

panies like SAP, Oracle, and Siebel offer mobile solutions that build on their existing enterprise systems offerings, primarily ERP, CRM, and SCM applications. These solutions address mobility in areas such as sales, field service, procurement, supply chain, and asset management. Mobile applications can be used to redesign or improve processes at a specific activity level (e.g., e-mail), or can be used to mobile enable large end-to-end processes that cut across functions (e.g., procure-to-pay processes). Similarly, mobile applications can be utilised for most processes in sales, supply chain, asset management, and plant maintenance. Systems that mobile enable core processes and key activities across multiple functions in an organisation can be seen as *mobile enterprise systems*. Mobile enterprise systems can either be enterprise systems extended to support process mobility or separate mobile systems integrated with existing enterprise systems. There are two key aspects to a *mobile enterprise system*. First, the mobile system should support one or more core business processes, and second, it works on the existing enterprise data.

As discussed earlier, the drivers for process mobility are location uncertainty and user mobility. As most sales and service staffs are highly mobile with activities carried out at external (uncertain) locations, these areas make a strong case for mobile systems. With mobile systems, the sales and field service staff can access business information wherever they like and capture data wherever it is generated. For example, sales personnel can view order status from a customer site and create new orders online using their mobile devices; service engineers can input job completion details on handheld computers that update centralised databases in real time. With the effective use of mobile systems, sales teams can spend more time with customers and prospects. Table 1

Table 1. Mobile enabling Sales and Service processes

CUSTOMER RELATIONSHIP MANAGEMENT (CRM)		
Sales: ▪ Product stock availability check ▪ Up-to-date product pricing information ▪ Delivery schedule ▪ View all customer information: Contacts, needs, decision makers, interaction history, etc.	Order Management: ▪ Order entry ▪ Order status ▪ Order delivery ▪ Payment collection	Account Management: ▪ View contact information ▪ View past interaction ▪ View planned events, meetings, etc. ▪ Negotiate and amend contracts in real time, printing these off, and e-mailing back to head office once agreed
Opportunity Management: ▪ View existing opportunities ▪ Capture new leads and opportunities	Activity & Task Management: ▪ View planned sales activities and tasks ▪ Create plan for new activities and tasks ▪ Capture customer signatures on a handheld device to confirm agreements	Field Service: ▪ Create service orders ▪ Capture assignment completion details ▪ Record customer complaints ▪ Manage service tasks and activities

Table 2. Mobile enabling Asset & Material management

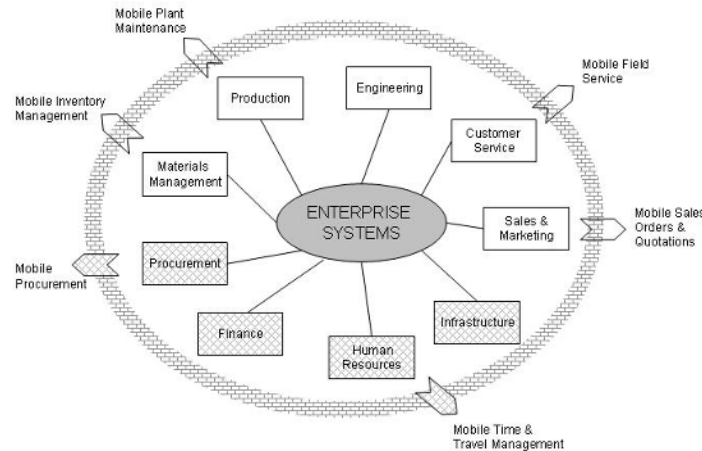
ASSET & MATERIALS MANAGEMENT	
Work Order Management: ▪ Work order creation ▪ Work order tracking ▪ Capture actual completion of tasks and activities ▪ Capture labour and material consumption ▪ Issue material against work order ▪ Raise requisition for material	Materials Management: ▪ Check material stock availability ▪ Issue and return material ▪ Physical stock count and balance adjustments
Equipment Management: ▪ View and update equipment repair history ▪ Record technical measurements of equipment ▪ Replace equipment at functional locations ▪ Manage asset hierarchy ▪ Install or uninstall equipment at functional locations	

lists some illustrative activities under sales and customer services that could be mobile enabled to improve process efficiencies.

Other high-potential areas are asset management, plant maintenance, and materials management. Mobility of workers in these processes is usually within a limited area, but requires movement around that area—for example, capturing technical measurements of equipment around the plant and updating equipment repair history on a handheld device. Mobile technologies, when combined with other

technologies such as bar code and more recently RFID (Radio Frequency Identification), can offer more appealing solutions and bring about substantial efficiency improvements. For instance, radio frequency (RF) tags can be used to store maintenance and service data pertaining to equipment. With the help of mobile devices, users can instantly view the equipment maintenance information and repair history stored on the RF tag attached on the equipment. Table 2 lists some of the activities that require

Figure 1. Mobile enterprise systems



local mobility and can obtain productivity gains from mobile enabling.

Supply chain management and procurement management processes can also benefit from mobile enabling. Mobile enabling employee-oriented processes such as filling and approving time sheets, travel expense forms, and leave requests can increase employee efficiency by making effective use of unproductive time such as travel by train, taxi, or air, as well as waiting periods.

With the growing use of mobile systems, mobile technologies are finding their place in enterprise systems' architecture and technology strategy. Organisations considering mobile systems for their core processes are viewing mobility as a strategic and not just technological element. From a systems perspective, mobile systems can be seen as a virtual mobile layer around the enterprise architecture. Similarly, the business process architecture can be understood to have a mobility layer that represents the mobility of business processes and supporting systems. Business organisations striving to be competitive will have to address mobility requirements and capture the opportunities arising from mobile enabling business

processes. Such organisations will require not only a mobile layer in their technology stack, but also a corporate-level strategy for mobility. Figure 1 shows an architectural view of the enterprise systems of an engineering business. The figure shows typical business areas using mobile systems as an extension of the enterprise systems. The mobile systems in the diagram are shown as a virtual layer around the enterprise systems.

REDESIGNING FOR MOBILITY

As mentioned earlier, mobile applications are now available for almost all key areas of business that require process mobility. With the proliferation of technology, the applications designed for mobile devices are nearing their desktop cousins in terms of functionality and performance. In order to leverage the capabilities of mobile systems, the design of business processes needs to be assessed from a mobility perspective and, if required, to be redesigned to maximise the advantage of mobile enabling. Mobile applications should not just be utilised to extend business processes onto handheld or

mobile devices. Instead, mobile systems should be tightly integrated with enterprise systems to streamline processes. Thus, technology innovation can be a stepping stone to achieve business process innovation. The ultimate objective of mobility in business can be seen as providing the power of a desktop computer in the hands of the mobile worker with the ability to connect to a network or the Internet from virtually anywhere. However, this may not be required or even feasible in all situations—technically or economically. Thus, it is important to identify and assess processes and aspects of processes that would benefit from mobile enabling to make a sound business case.

PROCESS REDESIGN PROJECTS

Most of the current interest and development in the area of business process change has its roots in the works done by Michael Hammer, James Champy, and Thomas Davenport in early 1990s. Since then the subject of business process change has evolved a great deal, with the practice growing across industries. Many different names have been given to such projects, for example, business transformation, business process improvement, and so on. Organisations change their processes with different objectives. All process change projects are different in a number of ways, mainly the objectives and the scope. Some projects have a strategic objective to achieve such as projects aiming to take advantage of strategic opportunities or address strategic problems. These include objectives such as to reduce cost, increase revenue, improve customer satisfaction, or strategically change business direction, while some projects have a specific goal of improving process efficiencies such as to implement an ERP/CRM system, Web-enable existing applications, or introduce electronic commerce. Projects can also vary in scope, as some cut

across multiple functions while others are focused on very specific sub-processes or activities.

Depending on the focus and objectives of the change, different terms are being employed to describe process change projects. In addition to that, some terms evolved simply due to poor usage of terms, which resulted in the terms being unfairly disliked; for example, the use of “Business Process Reengineering” has often been seen as a pseudonym for staff cutbacks. According to Harmon (2003, p. 39):

If the process is relatively stable and the goal is to introduce incremental improvements, then the preferred term is process improvement. If the process is very large and seeks to redesign the process in a comprehensive manner, then the term used is process reengineering.

Reengineering relies on re-conceptualising how the business process should work and is radical in nature. Process improvement is mostly done on a gradual and ongoing basis, and its objective is tactical problem solving. Reengineering is strategically oriented and aims to either seize strategic opportunities or address strategic business problems. The key difference is that there are no boundaries with reengineering while there are boundaries with process improvement. These two types of projects form the two extremes. For most process-change projects that fall in between these two extremes, the term generally used is *process redesign*. Most of the projects that aim to mobile enable processes will fall under process redesign. However, mobile enabling large processes can be part of a reengineering project, or mobile enabling a few activities can be an integral part of a process improvement initiative. For the sake of simplicity, I will use the term process redesign to refer to process changes that aim to address mobility needs.

PROCESS REDESIGN METHODOLOGY

In order to undertake a process change or process redesign project that involves mobile enabling processes, it is important to understand the considerations for redesigning to support mobility. A process redesign project may include a number of changes that help in achieving the process goals or overall project objectives such as: all orders should be processed in a maximum of two business days or reduce cost of end-to-end purchasing by 10%. For simplicity, I am only considering the mobility objective of a process redesign project. I am not considering changes done to achieve any other objectives. Regardless of the objectives of the redesign project, the overall approach to mobility should essentially be the same. Redesigning for mobility can be distinguished as a specialised pattern for redesign. According to Harmon (2003, p. 236), there are two sets of redesign patterns—namely, *basic business process redesign patterns* and *specialised business process redesign patterns*. Basic redesign patterns are generic approaches used by almost every process redesign project such as value-added analysis and reengineering. Specialised redesign patterns are used to extend the basic patterns or to solve specialised problems, such as ERP-driven design, workflow automation, and Six Sigma. Extending the patterns definition to mobility implies that another specialised redesign pattern can be defined, the driver for which is the need for mobility of business processes. I prefer the term “mobile enabling” for such a pattern.

The methodology applied depends on the type of project, scope of project, and individual preferences of the program managers. The activities involved vary depending on the methodology. The activities involved in basic methodology are explained briefly as follows:

- **Understand Business Goals:** As for any major redesign project, it is important to understand the overall objectives of the business, any strategic directions the business is considering, and how mobility ties into it.
- **Define Project Objectives and Scope:** The objectives and constraints of the project should be clearly defined and documented, for example, reduce the average processing time for service request from four days to three days. The definition should clearly explain how the strategic business objectives would be achieved or enabled by mobile enabling processes. The scope of the project should clearly define processes that will be covered, define the boundary of the scope by explicitly describing out-of-scope elements, and define the activities that will be covered by internal teams and those that will be contracted to outside vendors or consultants. The project charter should also include the benefits expected after the successful completion of the mobile-enabling project.
- **Define Project Plan:** The project plan should include the time, effort, and resources required for each activity involved. The inputs and outputs of each activity should be clearly identified and defined.
- **Analyse Existing Processes, People, and Technology:** The existing process are analysed to understand problems, gaps, disconnects, and any improvement areas. As part of the process analysis, the roles of the people involved need to be understood. In addition, the current use of technology in the processes needs to be clearly understood.
- **Evaluate Redesign and Technology Options:** Explore the possible redesign options and technology options. The redesign options are evaluated in light of project objectives, constraints, benefits, and costs.

For instance, for service engineers completing jobs at customer sites, do they need handheld computers to feed job completion data online into the enterprise database server, or can they make a phone call to a call centre and provide information which can then be input into the system. The evaluation in such cases is primarily driven by cost-benefit analysis. When it comes to personal productivity, in most cases providing access to wireless e-mails and a calendar can result in dramatic productivity improvements.

- **Redesign Processes, People Roles, and Technological Architecture:** Based on the most suitable option, new designs of processes job roles and systems to be implemented are established.
- **Define Implementation Plan:** The implementation plan includes the plan for introducing all the changes in processes, people, and technology. The implementation plan answers detailed and specific questions such as: Who owns the new field service process? and What system changes will be done? It includes the detailed plans for technical design and developments, testing, and so on.
- **Define Transition Plan:** The transition plan ensures that the transition to new processes and systems is controlled and smooth. The plan should address cutovers, systems failures, back-ups, and so forth.

In the following sections, I have included detailed explanations on two key aspects of redesigning processes for mobility.

ANALYSING EXISTING PROCESSES

When redesigning processes for mobility, there are two key activities: analysing current pro-

cesses and deriving the “to be” processes. In the analysis phase, analysing the distributed structure of a business process and identifying mobile sub-processes or activities assesses the need for mobility. Gruhn and Kohler (2003) suggest that when dealing with mobility of processes, we are dealing with a mobile partition or activity within a business process. A process can consist of one or more mobile activities, and as these activities affect the whole process, the complete business process is called *mobile business process*.

When analysing the distributed structure of the business processes in question, two critical questions need to be answered. First, where are the process activities to be executed, and second, what data is needed at which location of execution (Gruhn & Kohler, 2003). The other key consideration is the type of mobility of the user during the execution of the process activity. The different types of mobility of the worker executing the process activity can be classified into the following three modes (Kristoffersen & Ljungberg, 2000):

- **Wandering:** Worker performs activities while moving between different locations. The locations are locally defined within a building or local area.
- **Visiting:** Worker performs activities at different locations.
- **Travelling:** Worker performs activities while moving between different locations usually inside a vehicle.

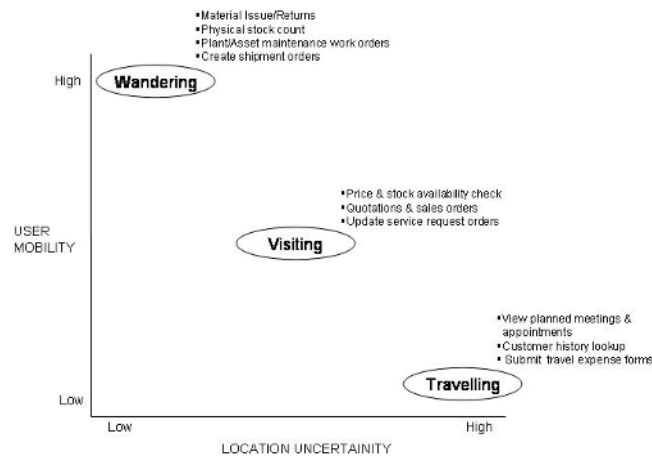
I will use the terms user mobility and worker mobility interchangeably to refer to mobility of the person executing the process activity. Valiente and van der Heijden (2002) defined the concept of “location uncertainty”, according to which the place of the execution of an activity can be different in different instances of the business process or the places can change during the execution of an activity.

The main characteristic of a mobile business process is that location uncertainty and worker mobility are externally determined, for instance: (a) a field service engineer inquiring for/requesting a repair part, and (b) a salesperson inquiring of the latest price and stock availability from a customer location. In both these examples, the location of activity execution is externally determined and worker mobility is of the type “visiting”. In another example where a maintenance engineer is updating work order details during plant/asset maintenance, though location uncertainty is low due to limited geographical area, it is externally determined. Worker mobility is of the type “wandering” in this case. Location uncertainty and user mobility are the chief determinants of the choice of mobile connectivity (e.g., GPRS, WLAN, Bluetooth, etc.) and type of mobile solution required. A sound mobile solution will appropriately address these two aspects of the mobile business process. Some sub-processes or activities could be performed when a user is mobile—that is, travelling or away from the office—as location is irrelevant for their execution. Leveraging location freedom can help increase productivity. Business processes that include such activities or sub-processes can be termed mobile friendly processes. Time and travel management are typical mobile friendly processes (e.g., filling and submitting time sheets, travel expense forms, and leave requests). Similarly, a number of workflow activities such as approving requisitions and leave requests are also mobile friendly. These activities can be performed through a mobile system when the user is travelling or away from the office and make substantial productivity gains for the individual. This may not be true in all cases, as all productivity gains do not always lead to real productivity improvements. It is critical to note however that in certain roles

this may not lead to a productivity gain for the company; for example, workers who do unpaid overtime will frequently utilise overtime to perform time management activities, and therefore there may be no productivity gain. However, even in such cases the business can gain other benefits such as reduced transport costs, as the staff is not required to travel to the office for administrative purposes. Horizontal processes, such as those discussed here, are generally mobile friendly but do not directly add value to customer-oriented processes. They can however add indirect value in terms of productivity gains and more time for customer-oriented processes.

The activities in the business processes of an enterprise can be segregated based on location uncertainty and worker mobility. Figure 2 shows some typical processes in a large engineering business. The processes are mapped against varying degrees of location uncertainty and user mobility (Valiente & van der Heijden, 2002). This is useful for understanding the drivers of mobility and as a guiding tool for classifying processes for mobile enabling. After identifying the processes that need to be mobile enabled, managers then need to consider the value that mobile systems would add to these business processes. They need to consider how well the process goals are achieved and how best the processes can be streamlined with the support of mobile systems. The identified processes can then be prioritised based on the added business value. As part of the exercise, problems and bottlenecks in the processes are also analysed. For each of the business processes to be mobile enabled, the key is identifying mobile activities that need to be supported by mobile systems. The basic requirement to conceptualise a mobile system is to understand the location at which the mobile activity is to be performed, data required for the activity, and mode of user mobility. A more

Figure 2. Process mobility map



detailed analysis of activities can be performed to determine time taken to complete the activity and so on. I will focus only on specific areas of analysis that are relevant to mobility.

To summarise, the analysis of existing processes revolves around the following questions. The answers to these questions provide crucial direction in identifying candidate processes for mobile enabling.

- Is there any uncertainty of location of process execution?
- Which activities in the process have location uncertainty?
- What is the type of mobility of the person executing the specific activity?
- Will mobile enabling the process improve process efficiency or productivity of users or not?
- What is the estimated actual cost, revenue, or strategic impact on the business as a whole of enabling the user to work from multiple locations?

MODELLING EXISTING PROCESSES

In order to analyse the identified processes in further detail, process flow diagrams are drawn for the current processes, to show the details of activities and their flow in the process. Process diagrams use the concept of swimlanes to indicate the user, function, department, or technology responsible for the process or activity. From mobility perspective, we are interested in swimlanes drawn on process diagrams depicting activity level detail. Swimlanes are also known as actor lanes, as the lanes can be used to represent the worker who performs the activities in the respective lane. Swimlanes can be depicted as either vertical lanes or horizontal lanes. Technical analysts and authors tend to prefer vertical swimlanes due to their similarity to UML, while those from an organisational or pure process background tend to prefer horizontal; either can be reviewed so it can be left as a matter for individual companies to decide. Figure 3 is a simplified example of a horizontal

Figure 3. Conventional process diagram

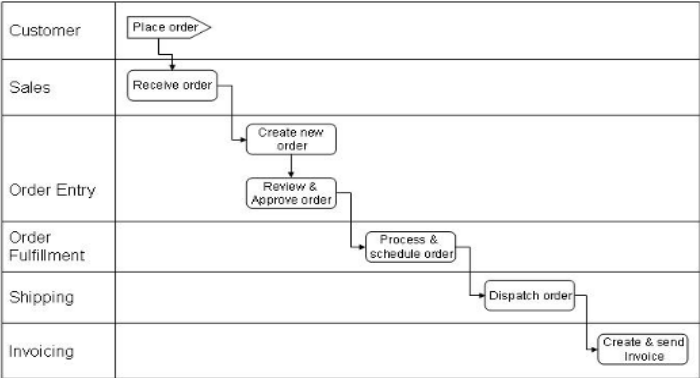
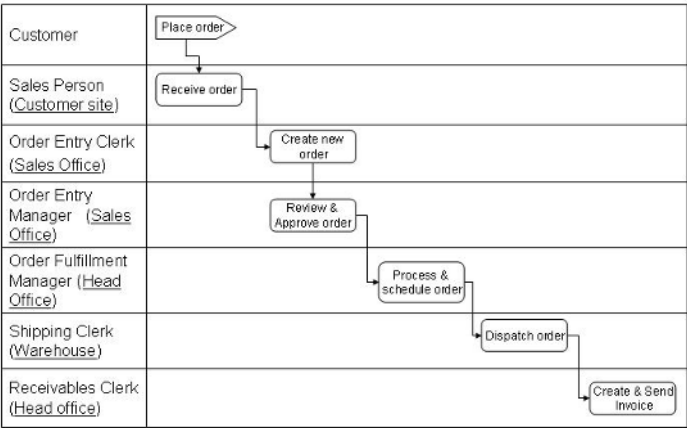


Figure 4. Process diagram with location information



swimlane for the sales process in an engineering company.

Conventionally, locations are not specified on the swimlanes, as they are not relevant to modelling processes for traditional information systems. However, from a mobility perspective, location is key. In order to reflect the location aspect on process diagrams, swimlanes can be shown for a user and location combination (Valiente & van der Heijden, 2002). Thus the labels of lanes should indicate location along with the actor; such process diagrams can provide valuable insight into mobility needs. It is

quite possible that a user is at more than one location to perform his or her activities, in which case that user can appear on multiple lanes, representing the multiple locations where the process activity is performed. However, there can be only be one lane for the required user-location combination, on the process diagram. These diagrams can highlight activities that are performed at an externally determined location such as customer site. The diagrams can be enhanced to capture high-level details of the data required for the mobile activities. Figure 4 shows a simplified example of process diagram

with mobility information for the same sales order process. Such a diagram can help identify the activities for potential mobile enabling and role changes; for example, the salesperson at the customer site can do the order entry instead of the order entry clerk doing it at the sales office.

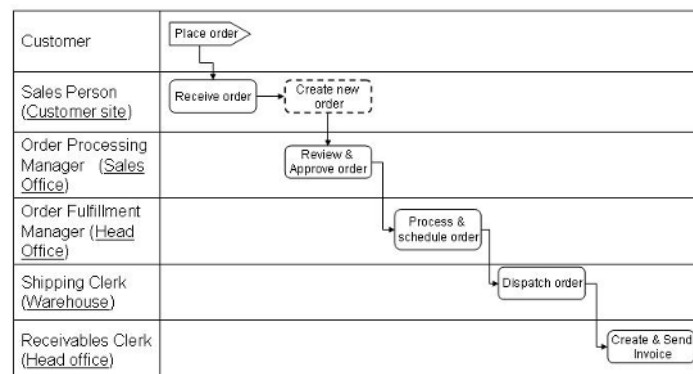
DESIGNING “TO BE” PROCESSES

After the detailed analysis of the current processes, various redesign options are explored. There are more considerations and complexities involved as I move on to the next stage of redesigning mobile business processes. In order to arrive at the best redesign option, the objectives of mobility need to be clearly defined in order to drive the options at this stage. After the analysis is complete, it is useful to review the technology options and constraints at a broad level along with the generic redesign options. It is at this stage that new ideas for redesigning processes are generated and options established. The objectives can be increased efficiency, improved realisation of process goals, or simply productivity gains. The best option in general is the one that best achieves these objectives. However, there can

be a trade-off between the cost and benefits of the options. In such situations, the most suitable option is the one that adds maximum business value. In some situations, a detailed cost-benefit analysis is also conducted to decide on the best option. The new design is captured in “to be” process diagrams, as this represents the design “to be” implemented. “To be” process diagrams should be drawn in a similar way as the existing process diagrams. The “to be” process diagrams can provide necessary inputs to define high-level requirements of the mobile element of the information system. Figure 5 illustrates a process diagram for a process that is redesigned to support mobility.

There are additional considerations for defining requirements for mobile systems compared to that for traditional information systems. Some important ones are: (a) understanding which mobile systems model is to be used, (b) controlling the access of enterprise data on mobile device, and (c) the profile of the users. The user is really at the centre of the design. The mobile systems model can be understood as a basic framework for the design of the mobile information system. The model determines how the data in the enterprise systems will be mobile enabled to support process mobility, as well as the application design and deployment of the mobile system.

Figure 5. “To be” process redesigned for mobility



There are three basic mobile systems models that have evolved, based on which enterprise mobile systems can be conceptualised.

1. Build a new mobile application using existing enterprise data
2. Build a cut-down “mobile” version of the current application
3. Entire existing application is mobile enabled to run on mobile devices

The choice of model(s) is determined by a number of factors, including the type of users to be supported, the type of mobility to be supported, the mobile devices to be supported such as laptops and handhelds, the capabilities of mobile connectivity, and the target time for activity completion. Organisations can either use one of the models or any combination of the three models depending on their needs or constraints if any. In general, the first two models are preferable for a number of reasons. The primary reason is that hardware and connectivity requirements for desktop applications are very high and cannot be met on expensive and compact mobile devices with limited screen sizes. The other reason is the advantage of reduced data entry time on the mobile version of the application, which is pivotal to the efficiency of a mobile worker. Using the first two models is also an opportunity to address the need of users who generally prefer to deal with a minimal set of data for their system tasks. A separate “mobile” version of the application created for mobile users can fill this gap. In situations where large contracts have to be read and amended online, it becomes imperative to extend the existing application to laptops to provide larger screen sizes.

ENTERPRISE MOBILITY STRATEGY

Businesses considering enterprise-wide process mobility will require a mobility strategy. The mobility strategy should guide operations and technology employees through the process redesign, application design, and implementation of the mobile systems. Mobility strategies will depend on a number of factors, including the nature of the business, its strategic goals, need for process mobility, existing IT investments, and financial budgets. As a result, each organisation’s mobility strategy will be unique in the same way their IT strategy is unique; one thing all sound mobility strategies will have in common, however, is that they will leverage existing IT investments and ensure maximisation of the return on investment on mobility projects.

There are a number of common factors that can be considered to devise an effective mobility strategy. I have identified a few key considerations for formulating an effective mobility strategy, which are as follows:

- **User Profiles:** Within each organisation there will be multiple “mobile worker” profiles, as the needs of mobile workers with different roles are likely to be very different from each other. Some mobile workers may require complex applications on a laptop, while others may need basic applications on handheld devices or a PDA.
- **Interoperability of Solution and Choice of Devices:** An effective mobile solution will be platform independent. The solution chosen should support multiple mobile devices and protect existing investments in IT infrastructure. The mobility strategy should clearly define the types of mobile

devices that will be supported at different stages, such as handhelds and PDAs. The mobile devices to be used should be carefully selected in line with the applications and technologies to be supported and the needs of the users. For users who require reading large reports and contracts, the devices should have larger screen sizes.

- **Mobile Infrastructure:** The mobile infrastructure chosen should allow applications to be accessed by multiple channels and device types so that the implementation teams can focus on functional and operational issues, and not get bogged down with deployment complexities. Setting up and maintaining a stand-alone mobile infrastructure can be very costly. Thus, extending the existing infrastructure is a cost-effective approach. Mobile infrastructure and applications chosen should aim to leverage existing investments in IT infrastructure and enterprise systems.
- **Mobile Connectivity:** The choice of mobile connectivity is driven primarily by the type of mobile application and the geography to be supported. The connectivity medium chosen should be scalable for the future needs of existing application or newer mobile applications. If a combination of mobile connectivity options is to be used (e.g., Wi-Fi and GSM networks), it is critical to understand the hand-off capabilities of the service provider or the technology deployed. In addition to the above, when choosing mobile connectivity it is important to consider the regional and global support of the chosen standard depending on geographies covered by the business; for example, GPRS is more global than CDMA, which only works in a few countries like the U.S., Canada, Japan, and Korea.
- **Service Providers:** It is also critical to assess the selected solution providers in terms of their existing and future capabilities of products and services. The requirements should be clearly mapped with vendor capabilities, and the capabilities of the chosen solution providers should be scalable in line with the mobility strategy of the organisation.
- **Alignment with Business Strategy:** The mobility strategy should be aligned with the business strategy. In other words, the business strategy should drive the mobility strategy, and not the other way around. If, for example, the business strategy is to reduce customer service costs, then in such a scenario, mobile enabling customer service processes will take priority over the sales processes.
- **Maximise ROI:** Mobile enabling business processes could involve major investments, and such investments have to be justified in terms of their return. As a guiding principle, businesses should first consider mobile enabling processes that provide maximum return on investment on mobile enabling. In simple terms, processes that can provide maximum cost savings or new revenue generation potential should be top on the list.
- **Beginner or Experienced Organisation:** If an organisation is experimenting with mobile systems for the first time, then it is rational to begin with horizontal processes instead of vertical processes. Horizontal processes such as time management are internal processes that are low risk and span across functions and departments. The advantage in doing so is that there is wider acceptance of new systems within the organisation, and the return on investment is faster.

- **User Requirements:** Mobility strategies will need to include direction for providing support for users in using new mobile technologies and tools. Additional considerations included here could be maximum weight of combined mobile articles per person, insurance, and the equipment the user should have access to within the office (e.g., should a user with a laptop also be allocated a desktop?).
- **Systems Security:** As mobile systems have more risks than conventional systems, it is vital to have a mobile systems security strategy in place. For example, theft or loss of a handheld device or unauthorised access to wireless networks can pose a serious threat to critical enterprise data. The mobile security strategy should be a part of the IT security strategy as well as the enterprise mobility strategy. The mobile security strategy should determine the type and extent of data that can be accessed on a combination of mobile devices and user profiles.

BUSINESS CASE FOR MOBILE ENABLING PROCESSES

Organisations have to make significant investments in order to mobile enable business processes and implement mobile systems. In order to make investment decisions, decision makers need a case for investment supported by cost-benefit analysis, so managers who propose to mobile enable their processes should provide a detailed business case to justify the investments. The business case should provide an assessment of the value of benefits in comparison to the total costs incurred in achieving process mobility. Quantifying the monetary value of benefits, particularly intangible benefits such as improved customer satisfaction,

can be more complex. There are number of books published that discuss business case development and cost-benefit analysis in detail which can be used to assist less experienced analysts in this area. I have enlisted some of the main benefits of mobile enabling business processes.

- Increased speed of end-to-end process execution and increased accuracy
- Enhanced reputation of customer services and increased customer satisfaction
- Reduced costs such as inventory, travel, and other administrative activities
- Enhanced decision making
- Increased productivity

A more scientific approach to measuring those benefits that cannot be measured easily is process metrics or measurements, and these are frequently used as part of standard process management practice. Process metrics, also known as KPIs (Key Performance Indicators), provide insight into the performance of the processes and hence can be utilised to measure the benefits of mobile enabling business processes. Metrics used will vary from industry to industry and from business to business dependent on each organisation's strategy. These metrics can help assess the improvements in process execution and the value added thereof.

Some illustrative process measurements for an engineering services business are:

- Average order fulfilment time
- Average service request resolution time
- Timeliness and accuracy of delivery
- Percentage sales returns
- Customer satisfaction ratings
- Average service down time

Process measurements are gathered initially before the processes are changed and

new systems are implemented. These measurements are captured again after the processes have been changed and new systems are put to use. The variance in measurements indicates the improvements in the process and can be attributed to the process and system changes. Gathering process measurements can be a very involved task, and it is recommended that initially these be kept to a small set of indicators until overall business value can be proven to prevent excessive expenditure on measurement, reducing the value of the actual improvements made. Organisations that collect process measurements on a regular basis are well placed to measure the benefits of the mobile-enabling processes. Process redesign projects aimed at mobile enabling business processes may also undertake some redesign that is not related to mobility. In such cases, the process measurements may not provide an accurate picture of benefits realised, due solely to mobility of processes if it is not possible to distinguish between benefits derived from mobility and those derived from other process changes. In this case, it is recommended that managers assign a proportion of improvement in measurements to each aspect of the process improvement including mobility.

REFERENCES

- Computer Business Review Online. (2005a). *Free enterprise*. Retrieved January 22, 2005, from www.cbronline.com
- Computer Business Review Online. (2005b). *Going mobile*. Retrieved January 22, 2005, from www.cbronline.com
- Davenport, T.H. (1993). *Process innovation: Reengineering work through information technology*. Boston: Harvard Business School Press.
- Gruhn, V., & Kohler, A. (2003). *Analysis of mobile business processes for the design of mobile information systems* (pp. 1-5). Chair of Applied Telematics/E-Business, University of Leipzig, Germany.
- Harmon, P. (2003). *Business process change—A manager's guide to improving, redesigning and automating processes*. San Francisco: Morgan Kaufmann.
- Henley, J. (2004). *Building the case for mobile business* (pp. 1-2). Redwood Shores, CA: Oracle Corporation.
- IT Facts. (2005). Retrieved April 2005 from www.itfacts.biz
- Mobile Advisor Magazine Online. (2005). Retrieved March/April 2005 from www.advisor.com
- Pak, C. (2004). *Evaluating the business value of mobile enterprise systems* (pp. 1-5).
- SAP-Global. (2005). *Mobile solutions*. Retrieved January 23, 2005, from www.sap.com
- Valiente, P., & van der Heijden, H. (2002). *A method to identify opportunities for mobile business processes* (pp. 1-10). SSE/EFI Working Paper Series in Business Administration, Stockholm School of Economics, Sweden.

Chapter XLII

Evaluation of Mobile Technologies in the Context of Their Applications, Limitations, and Transformation

Abbass Ghanbary

University of Western Sydney, Australia

ABSTRACT

Emerging mobile technologies have changed the way we conduct business. This is because communication, more than anything else, has become extremely significant in the context of today's business. Organizations are looking for communication technologies and corresponding strategies to reach and serve their customers. Mobile technologies provide ability to communicate independent of time and location. Therefore, understanding mobile technologies and the process of transitioning the organization to a mobile organization is crucial to the success of adopting mobility in business. Such a process provides a robust basis for the organization's desire to reach a wide customer base. This chapter discusses the assessment of a business in the context of mobile technology, describes the application and limitations of mobile technology, presents a brief history of mobile technology and outlines an initial approach for transitioning an organization to a mobile organization.

INTRODUCTION

This chapter evaluates the effects of mobile technologies on business and outlines an initial process of transitioning to mobile business. In 1874, when Alexander Graham Bell invented the telephone, he could not have imagined the significant impact it would have on number of aspects of human life. Similarly today, ad-

vancement in information and communication technology (ICT) has dramatically changed the way people live and conduct their businesses. One of the dramatic aspects of modern-day business activities is that these activities are conducted independent of location and time. For example, businesses are able to sell goods, facilitate customer enquiries, and coordinate their services through disparate geographical

and time boundaries primarily due to the wonders of communications technologies. Alter (1996) describes the ICT as tools for doing things, rather than just for monitoring performance of yesterday or last week. Thus it is quite logical to conclude that ICT has changed the very nature of the workplace.

The basis for the communications technologies in most modern business applications is the Internet. Increasingly, the required access and connection to the Internet has become very simple and ubiquitous in most developed nations. This Internet access has opened up opportunities for organizations to revolutionize their business processes. Undoubtedly, improvement of the communication technology has impacted not only our business domain, but also our socio-cultural domain. This, as per Unhelkar (2004), has resulted in the “next wave” of technologies called mobile technologies:

Mobile technologies are becoming the next technology wave as the increasing popularity and the functionality captures many hearts. Riding on the back of traditional Internet, mobile networks ensure that information is available to its users independent of a physical location.

This ubiquitous connectivity accorded by mobile networks referred to above impact has facilitated the increased communication between people. Furthermore, this mobile connectivity has also improved the ability of business processes to exchange data and conduct transactions.

This transformation of businesses has been evolutionary rather than revolutionary. For example, at the beginning of the Internet age, with the aid of its communications capabilities, businesses were transferred to e-business, and we even had the opportunity to do our daily business activities from home. Ghanbary (2003) has described the Internet as the most powerful

tool that brings information to our homes through communication lines, like water and electricity that come by power lines and pipes. Powerful search engines and the capability of sharing information are the great advantages of the Internet.

With the aforementioned strengths of mobile connectivity, it is also essential to work out a process that would outline “how” an organization can transition to such an m-enabled organization. However, this potential process of transitioning an organization to a mobile organization needs to incorporate all the major advances of mobile technologies of the past decade.

This is so because the philosophy of ordinary communication has given way to more advanced and efficient communications based on mobile and wireless technologies that enable business processes to be executed independent of time and location, resulting in a better, faster, and satisfactory response to the needs of the customer.

This impact of mobility is an important element of the mobile transition process that is felt at both business and personal levels. However, as of today, this process framework remains a challenge that needs to be further researched to enable businesses to transition successfully. This need for further investigation is also ratified by Ranjbar (2002), who correctly mentions: “It is not always possible to foresee all the implications of a new technology until it is adopted by the mass of population and used for a relatively long time.”

With the increase in the number of mobile organizations, the service providers realize that they need to identify their strengths as well as their weaknesses in terms of providing mobile services that provide solutions as well as rectify the shortcomings. The analyses of the weaknesses and strengths will give them an advantage to provide a convenient service and increase their customer loyalty.

BACKGROUND OF MOBILE TECHNOLOGY

The known mobile phones used today are the extension of American mobile radiotelephone. However, the distinction between such phones and two-way radios is not clearly known. The advancement on mobile technology and the relevant gadgets have been very moderate due to the limitation of this technology and the government regulations on radio transmission. The major concern of the American Communication Commission was to decide who would get what frequencies and which emergency service should have the priority on air for transmitting first.

The ordinary usage of mobile took place in the 1980s. The first generation of cellular mobile phones was used widely only by transmitting analogue signals. The large size and very high prices of mobile devices in addition to the cost of the calls are the well-known facts of this period.

In the 1990s, the second generation of mobile phones was available in the market using advanced digital technology. Very fast signals, cheaper calls and handsets, more reliable services, and the smaller handset devices are the characteristics of this period.

The uncontrollable growth of mobile technology has created new culture in the business world. The use of mobile and Internet technology has passed their boundaries to a business and social revolution. The new technology has capabilities of text, voice, and videoconferences using wireless devices as well as the ability to connect to the World Wide Web.

According to Connors and Connors (2004), the application and the use of mobile and Internet technologies are to organize people into various common interest groups. These groups can vary from harmless fun to serious military or political operations.

Information and communication technology has enlightened the business activities. The use of mobile devices is going to be another crucial factor to remain in the competitive market. Vaghijiani and Teoh (2005) explain this phenomenon of mobile technologies as an enormous opportunity for players up and down the value chain, from the device suppliers to carriers to the end users.

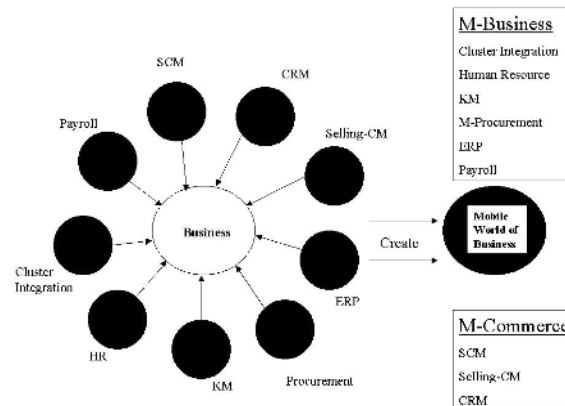
BACKGROUND TO A TRANSITION FRAMEWORK

New process frameworks need to build on existing work on process transitions. Electronic transitions have been studied and experimented by Ginige et al. (2002). In the electronic transitions, there has been ample focus on the effect of a dynamic environment and the rapidly evolving technology on organizations. Undoubtedly, these changes cause organizations to restructure and would introduce a new suite of business processes for them to enable them to remain in the market as well as grow by dealing with a greater number of customers.

The new business model and the use of technology in the development of these changes were the cause of the new term e-business. The term e-business might mean trade on the Internet for some managers, however it is looking at the facts in deeper methodology. The Australian e-business guide (Philipson, 2001) translates e-business as any business transaction or activity that uses the Internet. This includes not only the sale of goods and services directly over the Internet, but also the use of the Internet to promote and facilitate the sale of goods and services.

By using mobile phones or any other mobile devices, we are able to make our e-business model more accessible. The improvement and the efficiency will create more benefits, hence

Figure 1. Electronic mobile business model



the productivity remains even when people are out of their offices.

The proper design of the e-business model is a necessary component for the success of the m-business model. M-business makes the practice of the e-business model easier, more effective, and more profitable since there is no restriction on approaching the required data. The share of internal data, providing better and more reliable customer service and better control over the organization in general, are other major benefits of mobile business.

Figure 1 shows the re-engineered and mobilized individual enterprise enabling the transformation of an ordinary business to a new and modern world of mobile business. The business and commerce sections are given in different boxes that provide a more comprehensive study of the transformation.

Figure 1 represents electronic mobile business activities in more detail.

- **Clusters Integration:** Our mobile business model must be able to integrate all the clusters of the organizations. This fact might look easy, however by careful analysis it is realised that to connect all the clusters of the organizations is a huge task. It needs more than technological

advancement since some clusters might hesitate to share their information.

- **Customer Relationship Management:** The organizations must create close relationships with their customers. Customers want reliable and fast service. If the organizations provide them with what they want, they gain more business, and more business basically means more revenue. The CRM could be classified as a crucial factor of a mobile business, as it is in direct contact with people who are outside of the company. These people practically do not care how things are running as long as the great service is provided to them. A combination of technology, software, people, and re-engineered business processes are the fundamental of great customer relationship management.
- **Selling Change Management:** The right mobile business model should provide information about the available product directly to the customers. Direct interaction with the customers eliminates the retailer, and this will enable the business officials to provide detailed information about the product to the clients. As there are fewer hands involved in a purchase order, the prices offered to the customers will be dramatically low.
- **Enterprise Resource Planning:** The organizations have realised the importance of having the knowledge about back-office systems which could improve their customer order, integration of their clusters, and provide them with more sufficient information on how to run day-to-day activities by reducing the cost.
- **Supply Chain Management:** The technology is enabling the organizations to eliminate unnecessary processes to save time as well as money. The supply chain management of a business is the plan for

materials to be directed to the customers as quickly as possible by cutting the inessential retailers. The information about the delivery, financial matters, as well as order transmissions are provided on any mobile gadgets. The mobile business has more power of monitoring and control over the order status.

- **Procurement:** As per Kalakota (1999), purchasing refers to the actual buying of materials and those activities associated with the buying process. Procurement on the other hand has a broader meaning and includes purchasing, transportation, warehousing, and inbound receiving. Organizations spend millions of dollars for procurement every year. Mobile procurement gives them the opportunity to have a better control on their inventory, better control over the purchase approval, and so on. They have better control over the cost as well as knowledge regarding their assets.
- **Human Resources:** The functionality of human resources in the mobile world could be classified as providing the title of available jobs on mobile devices and providing more sufficient information about the positions and required level of desired criteria on the Internet. Enterprise Bargaining Agreement online and other responsibilities of HR could be done while using the new technologies.
- **Payroll:** The available technology is giving the opportunity to advance the payroll that saves a great deal of time and human resources. This strategy is more beneficial to the organizations with odd clusters such as shift workers. If their roster is also automated, the payroll office could automatically generate a payroll list based on the automated roster, while only a human supervisor is required.

- **Knowledge Management:** In today's competitive business world, knowledge plays the crucial rule. The knowledge must be reliable and accessible through all clusters of the organization anywhere and anytime. Knowledge created by the supplier is about available services and products, and knowledge about the users is in the form of profiles. Imagine if a customer calls the engineering department of his electricity company to get the approval for the extension of his house. At the end the inquiry, he asks for the amount of his bill; the system should be able to provide the necessary information rather than transferring his call to another operator. The available information should support the day-to-day running of the business.

By the aid of the emerging mobile technology and re-engineering the individual departments of the organization, the new mobile organization is created. In general, the purposes of the value-added services are to impress customers and create more control over the business. By replacing business with m-business, we can reduce the cost and create more revenue by having more satisfied customers.

APPLYING MOBILE TECHNOLOGIES

By correct application of mobile technologies into the business processes, the business enterprises are likely to gain advantages such as increased profits, satisfied customers, and greater customer loyalty. These customer-related advantages will accrue only when the organization investigates its customer behaviour in the context of the mobile environment.

In general, the application of mobile technologies could be classified in two different

categories, online and off-line services. The applications of online services are the executed applications when the mobile gadgets are connected to the mobile Internet.

The applications of the online wireless mobile Internet depends on situation (walking, driving), place (remote area, city metropolitan area), goal (aim of the connection), immediacy (instant action and reaction to demand), and load (how occupied the Internet provider is at the time of the connection). The congestion of the network clearly depends on the usage, which varies at different times of the day. As expected during the business hours, the network load on the Internet service provider is heavy.

The major online mobile applications are:

- **Information:** General information about movies. Location of cinemas, hotels, hospitals, and universities. News, sports, travel, weather, and financial information.
- **E-Mail:** To send and receive mail while online using the mobile handheld.
- **Payment:** To buy the product and receive the service and pay by your mobile device and receive the payment on your mobile bill. There is trusted third party in mobile commerce regarding billing inquiries that increase the cost since another party is involved.
- **Mobile Internet Banking:** To complete the banking transaction using your mobile device while online. The participating banks decide what kind of transaction is allowed and how they provide the security for their clients.
- **Mobile Internet Shopping:** To shop online using the mobile gadgets. There are advantages and disadvantages in mobile shopping since the participants are not able to touch or smell the items they are buying unless it is first sale vs. the repeated sale.
- **Education:** Using mobile handset to download lectures, use library facilities (order book, search the library), and use laboratory.
- **Government Applications:** Election, government bulletin and broadcasting, disaster information system with the aid of location-based services, and automatically giving the priority to the broadcast.
- **Communication:** Videoconferencing, telephony, sending and receiving pictures, and international communication.
- **Leisure:** Download music, video, TV, and games, and for some particular people gambling could be classified as a leisure activity.
- **Telemetric:** Location-based services, global positioning services, and car navigation systems. Telemetric applications could be expanded to give the opportunity to your device to book a hotel room, purchase a ticket, gather your required information, and any related scenario just by a click of the button or voice order, assuming your personal mobile gadget is already holding all your personal and credit card details. These functions could be performed by connecting to your mobile Internet or just by connecting to your network provider.
- **Advertising:** Conjunction of mobile and Internet technology for advertising. Receiving an advertisement on the Internet (pop-up screens) is not something new; furthermore marketing organizations could use the same idea for mobile Internet.

The off-line applications are the services offered by related network providers and extra available features on the particular mobile devices. Networks must have infrastructure to support the fast transmission of the data, reliability of the data, the integrity of the data, and

quality of service. The major off-line mobile applications are:

- **Communication:** Phone calls, SMS, messages, sending and receiving pictures.
- **Memory:** Phone book, music, different sounds, different effects (vibrate, volume), display, storing desired pictures and schedules and entertainment.
- **Expert System:** It would be possible for the organizations to use their mobile device as an expert system if their network provider has the capability to support it.
- **Remote Supervision:** To have control over the personnel while they are in an inaccessible area.
- **Traffic Information System:** Informing the drivers of traffic locations when they are driving close to the congested area.
- **M-Newspaper:** Subscribing to a newspaper if provided by the newspaper agency.
- **Advertising:** Based on the device's position, receiving the local advertisement.

Consumers' demands and corporate objectives could be different in the m-enabled world. While the application remains the same, expectation and usage are different. Usage in an m-enabled society is classified in three categories of interaction (voice, e-mail, chat, digital postcards, etc.), trading and business (banking, shopping, auctions, advertising, ticketing, etc.), and mobile-provided services (news, entertainment, driving direction, and much more).

It is very important to identify the sufficient information that is required to make the purchase decisions while the organizations are re-engineering their business processes. Ease of navigation and necessary links to other related Web sites are crucial factors for the software developers to consider while they are designing the new applications.

LIMITATION OF MOBILE TECHNOLOGY

Rising customer expectations have a direct connection to the advancement of technology. People's demand of the technology has not always been so realistic. The word "technology" has constantly fascinated human beings. Information and communication as the defining technology of the modern era have increased the expectations to an irrationally higher level. As Toffler (1980) predicted, people's dependence on technology has increased to a high level where technology has affected every aspect of human life. Mobile technology, which is an integration of communication and computer technology, has created such expectations in human behaviour that people cannot think of an era without such technology today.

It is clear that people rely on technology even when technology does not have the capability, or it is not robust enough, to support their task. There is no guarantee that I will not lose my work while writing this chapter on my computer, and the very same technology is used when human lives are involved. As an example, computers are used to take off and land airplanes.

However, mobile technology has its own characteristics and limitations which should be clearly identifiable to business enterprises. Of course these limitations will increase when mobile devices are connected to the mobile Internet.

Jamalipour (2003) explains that the access to the wireless mobile Internet is not just an extension of the Internet into the mobile environment giving users access to the Internet while on the move. However, it is about integrating the Internet and telecommunications technologies into a single system that covers all communication needs of people. He also believes that current network architectures used

in either the wired Internet or the cellular networks would not be appropriate and efficient for future wireless mobile Internet, even if we assume that the cellular network will provide the major infrastructure of the mobile Internet. He concludes by saying that access to the mobile Internet is slow, expensive, and confusing.

Some limitations of mobile technology are as follows:

- **Cost:** The cost of restructuring the organization and personal devices.
- **Call Drops:** Disconnection while taking or downloading the data.
- **Connectivity:** Constant connectivity to a network is a big issue for mobile network providers. There are improvements in this area, but it should be taken into consideration when we are mobilizing the enterprise that the network must support the expected assignment of the enterprise.
- **Lost Work:** Losing the performed work due to disconnection or dead battery.
- **Managing Technology:** Consistent maintenance of the software and the hardware.
- **Security:** Payment online, user behaviour, rules and hassles, mobile virus protection, file encryption, access control, and authentication are the most important security factors in the mobile environment, considering that mobile devices are very personal; in case of loss or theft, who is accessing the corporate or personal data?
- **Integrity:** The transmitted data is actually going to the expected individual. The message received is actually the message sent, and also the sender is the real owner of the mobile handset.
- **Privacy:** Who is accessing the corporate database and where personal details of the individuals are involved.
- **Regulations:** Government roles and regulations regarding the mobile matters.
- **Standardization:** Technical standards and compatibility of the users (business-to-business, business-to-customer).
- **Health Hazards:** By encouraging people to use mobile gadgets, are we jeopardizing their health?
- **Data Transmission Speed:** Slow transmission is very costly and ineffective.
- **Coverage:** The coverage of the network in a remote area is an identified and unresolved problem.
- **Adaptation:** Some people are resistant towards technology, and it would take time for them to adapt to the new technology.
- **Training:** The cost of training, managing mobile workforce, and controlling their activity.
- **Marketing Issues:** It would be a new era of marketing issues in mobile age such as sex, age, and so on.
- **Social Aspects:** Technology is creating a new pressure for ordinary people. People resistant to changes should be the major concern while planning the mobile transformation. Perpetual contact is another issue that is changing the face of our society. Mobile users are communicating with some other person while driving, walking on the street, and when they are in different public places. It is becoming commonly acceptable in our society to give priority to the mobile caller even when personal face-to-face conversation is getting disrupted.

Limited processing powers of handsets' microprocessors, memory size, battery life, small screen of handheld devices and their resolution, replacement costs, required ongoing support, network charges, as well as mobile Internet

charges and enhancements are the other critical shortcomings of the mobile technology.

TRANSITIONING TO A MOBILE ORGANIZATION

The transformation of the organization by introducing the new and re-engineered processes is a very crucial matter. Should the enterprise revolutionize and transform as soon as possible or use the evolutionary process? Should the enterprise adapt to the new technology as soon as it is available or delay the process to see the outcome by using another organization's experience?

Considering there is no suitable answer for the above questions, there may be a need for a new approach to clarify this uncertainty. This new approach is supposed to be the combination of the revolution and evolution—revolution since the organization should not fall behind by remaining competitive in the market, and evolution to reduce the risk of not having a successful e-transformation. According to Murugeson and Deshpande (2001), the development of an organization could be classified as the following:

The choice of a suitable development model, according to practitioners and researchers, is site (and applications), its document orientation, content and graphic design, budget and time constraints and the changing technology.

It could be quite risky if the organisations adapt to the new technology as soon as it is available in the market, as the system is definitely unknown and there might be hundreds of unresolved issues as well. Another factor at an early stage entry is the high cost involved in the introductory level of the new technology.

However, if they do not adapt during a specific period of time and their competitors do,

there is a great chance of not being able to catch up with the advancements in the professional world. These are some issues that management is facing today. Their crucial decision making will determine the future of their companies. Serour (2005) clarifies that senior and middle management find it hard to proceed when there is (still) very little guidance available from real-world experience.

The organizations must allow internal and external parties involved to know that there are some changes that need to take place. In view of people's resistance to change, this will give them some time to prepare and adjust. The core of the training is for internal parties of the organization; however it is very important to provide sufficient information to external parties and advise them about the change.

The organizations must plan and manage change (cultural, technological, internal, and external) and understand the key areas associated to dangers related to their working environment that others have discovered and faced. The Australian e-business guide (Philipson, 2001) describes that implementations for e-business initiatives must be rapid and each project should be delivered in a maximum of three months. Build quickly and move to the learning stage, then build the next stage and fix the previous ones based on what you have learned.

Management must support the variation in business and market strategies, organizational restructure, and management strategies. The corporations must prepare all the existing clusters ready for change. Managing the transformation by having a reliable and calculated plan is the crucial factor for success. The transition must remain persistent alongside with detailed knowledge of the development of the individual clusters. According to Brans (2003), generally mobile transition takes place by distinguishing what kind of portable devices, networks, appli-

cation gateways, and enterprise applications are required.

The benefits of mobilizing the organizations are: quick sale, closer communication within the internal departments, more strength and opportunities and less weaknesses and threats, professional façade for the organization, quick and reliable generation of customer data, and mobility at work.

CONCLUSION AND FUTURE DIRECTIONS

This chapter described some characteristics of m-business and offered a brief background of mobile phone technology.

When the Internet was introduced, nobody could imagine that this tool was going to make the next paradigm shift in all human interaction as well as business transactions. M-business is enabling organizations to increase global productivity. With the aid of mobile technology, the capability exists to operate in a very modern and extraordinary manner. The problems faced in the transformation of an organization to m-organization were identified and some solutions were recommended.

The domain of this chapter was to explain the particulars of a mobile business model, emphasising their significance, application, and the shortcomings of this technology in the world of business and trade. It is hoped that this chapter could convince the developers of the m-applications to spend more time in their design to fulfil the needs of the end users.

However, there are some critical issues that are unresolved in the world of mobile technology. These issues can be classified as security (integrity and privacy), national/international regulation, international standardization, security and integrity of databases on mobile devices, managing mobile workers and coordina-

tion of their activities, and the consistent maintenance of mobile hardware and software.

To create a robust and reliable mobile world, the developers could consider some other shortcomings of mobile technology. These issues are mobile payments, health hazards, the cost of restructuring the organization, damage to the handheld devices, legal liability of handheld devices (since the mobile gadgets are very personal and can keep confidential data related to the organization), and constant connectivity of mobile devices.

REFERENCES

- Alter, S. (1996). *Information systems. A management perspective*. Benjamin/Cummings.
- Brans, P. (2003). *Mobilize your enterprise*. Pearson Education.
- Connors, J., & Connors, S. (2004). The impact of mobile technology on business planning. *Proceedings of IRMA 2004*, New Orleans, LA.
- Deshpande, Y., & Ginige, A. (2001). Corporate Web development: From process infancy to maturity. In S. Murugesan, & Y. Deshpande (Eds.), *Web engineering managing diversity and complexity of Web application development* (p. 36). Germany: Springer-Verlag.
- Ghanbary, A. (2003). *Effects of computers on family and leisure time*. Honour Thesis, University of Western Sydney, Australia.
- Ginige, A. (2002). New paradigm for developing evolutionary software to support business. In S. K. Chang (Ed.), *Handbook of software engineering and knowledge engineering* (Vol. 2). World Scientific.
- Jamalipour, A. (2003). *Wireless mobile Internet: Architectures, protocols and ser-*

vices. Hoboken, NJ: John Wiley & Sons.

Kalakota, R., & Robinson, M. (1999). *E-business roadmap for success*. Boston: Addison Wesley Longman.

Murugesan, S., & Deshpande, Y. (2001). *Web engineering (publication data)*. Berlin/Heidelberg: Springer-Verlag.

Philipson, G. (2001). *Australian e-business guide*. McPherson's Printing Group.

Ranjbar, M. (2002). *Social aspects of information technology*. Sydney, Australia: University of Western Sydney.

Serour, M. K. (2005). The organizational transformation process to globalization. In Y. Lan (Ed.), *Global information society: Operating information systems in a dynamic global*

business environment. Hershey, PA: Idea Group Publishing.

Toffler, A. (1980). *The third wave*. William Morrow and Company.

Unhelkar, B. (2005). Web services and their impact in creating a domain shift in the process of globalization. In Y. Lan (Ed.), *Global information society: Operating information systems in a dynamic global business environment*. Hershey, PA: Idea Group Publishing.

Vaghjiani, K., & Teoh, J. (2005). Comprehensive impact of mobile technology on business. In Y. Lan (Ed.), *Global information society: Operating information systems in a dynamic global business environment*. Hershey, PA: Idea Group Publishing.

Chapter XLIII

Policy-Based Mobile Computing

S. Rajeev

PSG College of Technology, India

S. N. Sivanandam

PSG College of Technology, India

K. V. Sreenaath

PSG College of Technology, India

ABSTRACT

Mobile computing is associated with mobility of hardware, data and software in computer applications. With growing mobile users, dynamicity in catering of mobile services becomes an important issue. Policies define the overall behavior of the system. Policy based approaches are very dynamic in nature because the events are triggered dynamically through policies, thereby suiting mobile applications. Much of the existing architectures fail to address important issues such as dynamicity in providing service, Service Level provisioning, policy based QoS and security aspects in mobile systems. In this chapter we propose policy based architectures and test results catering to different needs of mobile computing

INTRODUCTION

Policies are rules that govern the overall functioning of the system. *Policy computing* is used in a variety of areas. *Mobile computing*, with its ever-expanding networks and ever-growing number of users, needs to effectively implement a policy-based approach to enhance data communication. This can result in increasing customer satisfaction as well as efficient mobile network management.

POLICY COMPUTING AND NEED FOR POLICY-BASED MOBILE COMPUTING

Policies in society and organizations are often captured and enforced as laws, rules, procedures, contracts, agreements, and memorandums. Policies are rules that govern the choices of system behavior. A policy is defined as “a definite goal, course or method of action to

guide and determine present and future decisions.” Security policies define what actions are permitted or not permitted, for what or for whom, and under what conditions. Management policies define what actions need to be carried out when specific events occur within a system or what resources must be allocated under specific conditions. They are widely used for the mobile user whose requirements are dynamic.

Policy-based computing is the art of using policy-based approaches for effective and efficient computing; it is widely used because of its dynamicity. Hence in areas such as mobile computing, policy computing can be effectively used.

Much of the existing network systems’ are configured statically (Fankhauser, Schweikert, & Plattner, 1999). In the present-day scenario, the number of mobile/wireless network users increases day by day. With the static systems being deployed, it is very difficult to achieve the needed dynamicity for mobile computing resulting from changing user base. In order to achieve efficient communication for fluctuating user base, policy-based systems need to be implemented in different areas of the existing wireless mobile network infrastructure.

POLICY IN MOBILE COMPUTING

Mobile computing is conducted by intermittently connected users who access network resources that need to escalate with increasing computing needs. Mobile computing has expanded the role of broadcast radio in data communication, and with increasing users, providing quality service becomes a challenging issue. The mobile users must be provided with the best possible service so that the service provider can stay in competition with peer service providers. In order for the best possible

service to be provided to the mobile users, there are certain criteria that should be met. They are:

- The quality of service should be guaranteed.
- There should be effective service-level agreement (SLA) between the mobile user and the service provider.
- Security should be foolproof.

With the existing system (without a policy-based approach), it becomes very difficult to achieve the mentioned criteria. It is very difficult to provide a guaranteed quality of service (QoS), which is also dynamic (not statically configured). Moreover SLA is a very static procedure. Because of the mobility and dynamicity of mobile networks, SLAs also must be made very dynamic. Similarly, security should also be made very dynamic and efficient. To overcome all these shortcomings of the existing system, *a policy-based approach should be used in mobile networks.*

Policy computing can be effectively implemented in mobile networks using policy compilers. Policies can be written in different ways. There are different languages for writing policies that are used for different purposes of specifying policies. In order that the “security policies” be specified, languages such as Trust Policy Language (TPL), LaSCO, and so forth are used. In a similar way, for specifying management-related policies, languages such as Ponder, Policy Maker, and so forth are used. Thus for different scopes of application of policies, specific languages are used.

Policy validation checks a solution’s conformance to the policy file. The actual process of policy validation has three primary stages. First, a node or hierarchy change event in Solution Explorer (such as add, drag, or delete event) begins the validation process. Then the

validation process maps items discovered in the solution (such as files, references, classes, or interface definitions) to a corresponding Template Description Language (TDL) policy ELEMENT node. Finally, for recognized ELEMENT nodes, the validation process checks the parent ELEMENT for policy compatibility with the child ELEMENT. When the policies are compatible, the validation process applies any ELEMENT-specific policy.

APPLICATIONS OF POLICY COMPUTING

Policy-based management is an over-arching technology for an automated management of networks (Lewis, 1996). Policy-based management is being adopted widely for different domains like quality of service, wireless networks, service-level agreement, virtual private networks (VPNs), network security, and IP address allocation. Therefore policy-based networking configures and controls the various operational characteristics of a network as a whole, providing the network operator with a simplified, logically centralized, and automated control over the entire network. In a wireless/mobile network, events are user and time based. These events are very dynamic in nature in order to provide the best service to the mobile users and also to maximize the profit of the service provider. But most of the existing systems are very static in nature. With the static systems being deployed, it is very difficult to achieve the needed dynamicity for mobile computing. In order to achieve this, policy-based systems need to be implemented in different areas of the existing wireless mobile network infrastructure such as QoS in wireless networks (especially differentiated networks), security, and SLAs.

Policy-Based Architecture for Security

Some of the key issues involved in providing services for wireless networks are (Sivanandam, Santosh Rao, Pradeep, & Rajeev, 2003):

1. **Bandwidth Cost:** Depending upon the number of users connected, the location (e.g., urban or remote), and the type of service (e.g., video, audio, etc.) being offered, dynamic allocation of bandwidth plays an important role.
2. **Limited Memory:** Today's wireless device places constraints on the amount of data that it can hold. Moreover, this limit depends upon the device being used and hence causes greater concern with low memory devices.
3. **Access Cost:** Optimizing the cost (Boertien, Janssen, & Middelkoop, 2001) of accessing and transferring data is more complex in wireless networks than in wired networks. If the number of servers used by a service or the number of services provided by an enterprise increases, then maintaining service consistency would turn out to be a cost factor in itself.
4. **Scalability Requirements:** These requirements force the service provider to think in terms of developing a solution that would support increasing and decreasing the number of services offered by the enterprise.

These constraints adversely affect the process of implementation of wireless/mobile services over the existing architecture. To overcome this, an identity management architecture for a wireless differentiated service schema that could be implemented using LDAP (light-weight directory access protocol) (Hodges &

Morgan, 2002), directory structures are constructed.

Policy Warehouse

The concept of differentiated services entitles the maintenance of a large amount of information pertaining to the user (e.g., user names, passwords, services registered, premium amount, etc.) and an efficient quick access mechanism to retrieve the relevant details. This overhead increases when it comes to wireless services. Here, the policy warehouse acts as the information backbone of the service provisioning system. The service provisioning engine contacts the policy warehouse whenever the service provider forwards to it a request from the user, after the right user has been authenticated.

1. **Id-Synch and P-Synch:** The synchronization of user identities and passwords pertaining to a single user is highly crucial in providing a hassle-free connection to services that require subscription to external back-end service providers. This architecture uses Id-Synch and P-Synch mechanisms for identity synchronization and password synchronization respectively.
2. **Meta Directories:** In general, service providers need to maintain a global user profile to uniquely identify a user over the various services provided to him. This information, which mainly comprises a collection of information pertaining to the user sign-on details of various services, is stored in meta directories. This profile makes it easy for the service provisions engine to authenticate the user. The architecture has a provision of compiling as well as retrieving meta directory information.

3. **LDAP Access Engine and Directory Structures:** Information pertaining to the user is stored in lightweight directory structures that can be retrieved using the LDAP. Directory structures are used to store user information because they provide a systematic mechanism for organizing data under a common head like user profiles, user services, user privileges, and so forth that are organized in a hierarchical manner on multiple workstations that are distributed over a network. This not only makes data retrieval fast, querying complexity less, and volume of data storage minimum, but it also makes easy implementation of policies that depend on the enterprise using the system.

The LDAP engine acts as an interface between the various servers like Id-Synch and P-Synch, and directories like meta directories and the underlying LDAP directory structures. They process the request for data from the higher layers and hand over the appropriate data to the requested application in the required format.

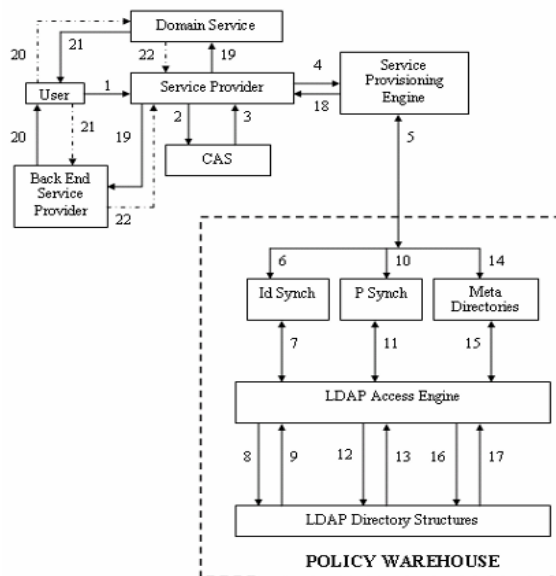
The LDAP directory services are part of the directory-enabled network services (DEN) that provide standard APIs for the access of network objects.

Service Provisioning Mechanism

The service provisioning system can be generally viewed in the following stages:

- **User Login:** In this stage, the user sends his details like User Name, Password, and Chip Index Number to the service provider for authentication.
- **Service Provider:** The Service Provider has to perform the following two actions on a service request:

Figure 1. Policy-based provisioning



1. **Authentication:** The authentication is accomplished using the DSAP (distributed substring authentication protocol) (Sivanandam et al., 2003). This is done by fragmenting the user details into sub-strings and distributing them over a network which is monitored by a central authentication system. When the user is required to authenticate, the protocol fetches the appropriate sub-strings from the network and compares them to the user input. A match signifies a valid user. After this stage, the appropriate user policy is fetched from the policy warehouse using the service provisioning engine.
2. **Providing the Service:** This is the last step of the service provisioning. In this stage, the actual service that the user has requested is granted. The service could be from a back-end service provider or from the main service provider. This detail is an

abstraction to the user who undertakes all transactions with the main service provider only. When the user disconnects from the service, intimation is sent to both the back-end service provider and to the main service provider. This has two implications: firstly, the main service provider's load is shared by the back-end service provider, and secondly, the intimation during the connection termination ensures that the main service provider gets the appropriate usage details. This can act as verification of the details that the back-end service provider will submit later.

Policy Based Architecture for QoS

The interface to the network device and the information models required for specifying policies are either standardized or being standardized in IETF and DMTF. An architecture for a policy-based QoS management system for Diffserv-based wireless networks, which are based on COPS for interfacing with the network device and on LDAP for interfacing with a directory server for storing policies, is constructed. The Diffserv policies are installed based on role combination assigned to the network device interfaces. The directory access could become a bottleneck in scaling the performance of the policy server, and it can be improved substantially by employing appropriate policy caching mechanisms. The framework considers various QoS parameters in the wireless network and proposes the policy-based architecture for QoS management in wireless networks.

Wireless Network QoS Parameters

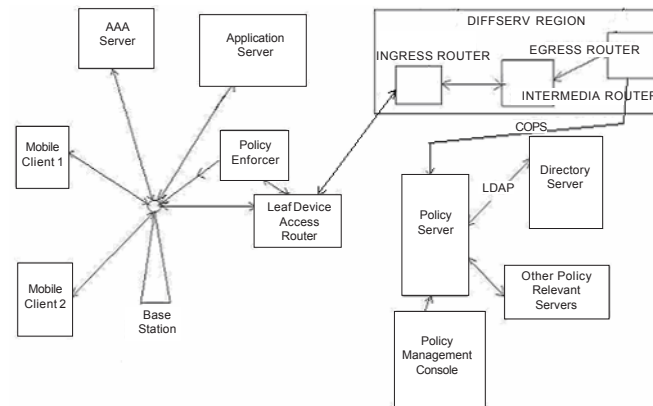
The wireless/mobile network is affected by the following QoS parameters:

- **High Loss Rate:** Wireless/mobile networks are characterized by more frequent packet losses because of fading effects. The scheduler may think that a certain DSCP is being satisfied with the required number of packets scheduled, but the receiver is not receiving the packets at the required rate. It will be useful to have feedback from the receiver so that some compensation techniques can be employed. The base station (BS) can better handle compensation of lost bandwidth using this information.
- **Battery Power Constraints:** Current mobile battery technology does not allow more than a few hours of continuous mobile operation. Two of the major consumers of power in a mobile network are the network interface (14%) and the CPU/memory (21%). Therefore, network protocols should be designed to be more energy efficient (Agrawal, Chen, & Sivalingam, 1999). The mobile device can use the signaling mechanism to periodically send messages about its power level to the BS. The BS can then use this information to dynamically decide packet scheduling, packet dropping, and so forth.
- **Classification of Packets within a Flow:** Present Diffserv (Chan, Sahita, Hahn, & McCloghrie, 2003) mechanisms treat all packets within a flow identically. Even though a distinction can be made between packets as in-profile or out-of-profile, all in-profile packets are treated the same way. In many situations (e.g., while using layered video), it may be necessary to distinguish packets within a flow. This is because some packets from a flow level could be more important than the others, and a local condition like power level may lead to different treatments of these packets. Thus, the packets within a flow must be made distinguishable, and bits in the TOS field may be used for this purpose. To summarize, the various possible factors needed to make the Diffserv architecture suitable for wireless networks were discussed in this section.
- **Low Bandwidth:** Wireless networks available today are mostly low bandwidth systems. Most of the current LANs operate at 2 Mbps with migration up to 11 Mbps available. However, the available wireless LAN bandwidth is still an order of magnitude less than the typical wired LAN bandwidth of 100 Mbps. This leads to two decisions. First, the signaling protocol should be very simple and highly scalable. It is also better to modify an existing protocol for compatibility with other existing network protocols. Second, the mobile should not be swamped with too much data from a wired sender with higher network bandwidth. This can be handled to a large extent by transport protocol control, but the problem can be alleviated by handling it partially at the base station. Therefore, mechanisms may be used at the BS to send data to the mobile devices based on current conditions such as channel condition, bandwidth available, and so forth.

Policy-Based QoS

The IETF Resource Allocation Protocol (RAP) working group has defined, among other standards, the policy-based admission control framework, and the common open policy service (COPS) protocol and its extension—COPS for provisioning (COPS-PR). COPS is a simple query protocol that facilitates communication between the policy clients and remote policy server(s). Two policy control models have been defined: outsourcing and provisioning. While COPS supports the outsourcing model, its ex-

Figure 2. Policy-based management system architecture



tension COPS-PR integrates both the outsourcing and provisioning models. The outsourcing model is tailored to signaling protocols such as the resource reservation protocol (RSVP) (Braden, Zhang, Berson, Herzog, & Jamin, 1997), which requires traffic management on a per-flow basis. On the other hand, the provisioning or configuration model is used to control aggregate traffic-handling mechanisms such as the Differentiated Services (Diffserv) architecture. In the outsourcing model, when the PEP receives an event (e.g., RSVP reservation request) that requires a new policy decision, it sends a request (REQ) message to the remote policy decision point (PDP). The PDP then makes a decision and sends a decision (DEC) message (e.g., accept or reject) back to the PEP. The outsourcing model is thus PEP driven and involves a direct 1:1 relation between PEP events and PDP decisions. On the other hand, the provisioning or configurations model (Chan et al., 2001) makes no assumptions of such direct one-to-one correlation between PEP events and PDP decisions. The PDP may proactively provision the PEP reacting to external events, PEP events, and any combination thereof (N: M correlation). Provisioning thus tends to be PDP driven and may be

performed in bulk (e.g., entire router QoS configuration) or in portions (e.g., updating a Diffserv marking filter).

Architecture of a Policy-Based Management System for a Diffserv-Based Wireless Network

Figure 2 illustrates the architecture of the policy-based management system for Diffserv-based wireless networks. The policy server is responsible for interpreting higher-level policies and translating them into device-specific commands for realizing those policies. For allocating resources on inter-domain links and for implementing SLAs, the policy server (especially the bandwidth broker component) has to communicate with the policy server in the provider.

The policy server is mainly responsible for the following:

- retrieving relevant policies created by the network administrator through the policy console after resolving any conflicts with existing policies;
- translating the policies relevant for each PEP into the corresponding policy information base (PIB) commands;

- arriving at policy decisions from relevant policies for policy decision requests, and maintaining those decision states; and
- taking appropriate actions such as deletion of existing decision states or modification of installed traffic control parameters in the PEP for any modifications to currently installed policies.

All the policies are stored in the LDAP server. The policy editor (PE) is the entity responsible for creating, modifying, or deleting policy rules or entries in the LDAP server. LDAP protocol provides access to directories supporting the X.500 models, while not incurring the resource requirements of the X.500 directory access protocol (DAP). It is specifically targeted at management applications and browser applications that provide read/write interactive access to directories. It does not have the mechanism to notify policy consumers of changes in the LDAP server. Therefore, it is the responsibility of the policy editor to indicate the changes in the LDAP server, as and when required, using an internal event messaging service. The policy server, in addition to querying the LDAP server, queries other policy-relevant servers such as Certificate server, Time server, and so on.

The policy management client—also referred to as the policy editor—provides a high-level user interface, for operator input translates this input into the proper schema for storage in the directory server and pushes it out to the directory for storage. The authentication, authorization, and accounting (AAA) server is responsible for authentication, authorization, and accounting of the user after the relevant policies have been picked and enforced in the policy enforcers (routers). This AAA server is used by the base station to check if the user is authenticated and authorized for the resource he requests, and to check if he is accounted.

The policy enforcer nearer to the base station enforces the policy decisions taken from the policy server. The base station then requests the nearest application server (after policies are enforced) and waits for the response from the application server.

The base station first sends the request to the leaf access router, which then sends it to the ingress router in the region. The ingress router then passes on the requests to the intermediate router. The request passes through the other intermediate routers and reaches the egress router, which sends the request to the policy server through COPS.

Policy-Based Architecture for SLA

Mobile ad hoc networks (MANETs) are autonomous networks operating either in isolation or as “stub networks” connecting to a fixed infrastructure. Depending on the nodes’ geographical positions, transceiver coverage patterns, transmission power levels, and co-channel interference levels, a network can be formed and unformed on the fly. Ad hoc networks have found a growing number of applications: wearable computing, disaster management/relief and other emergency operations, rapidly deployable military battle-site networks, and sensor fields, to name a few. The main characteristics of ad hoc networks are:

- **Dynamic Topological Changes:** Nodes are free to move about arbitrarily. Thus, the network topology may change randomly and rapidly over unpredictable times.
- **Bandwidth Constraints:** Wireless links have significantly lower capacity than wired links. Due to the effects such as multiple accesses, multi-path fading, noise, and signal interference, the capacity of a wireless link can be degraded over time and

the effective throughput may be less than the radio's maximum transmission capacity.

- **Multi-Hop Communications:** Due to signal propagation characteristics of wireless transceivers, ad hoc networks require the support of multi-hop communications; that is, mobile nodes that cannot reach the destination node directly will need to relay their messages through other nodes.
- **Limited Security:** Mobile wireless networks are generally more vulnerable to security threats than wired networks. The increased possibility of eavesdropping, spoofing, and denial-of-service (DoS) attacks should be carefully considered when an ad hoc wireless network system is designed.
- **Energy Constrained Nodes:** Mobile nodes rely on batteries for proper operation. As an ad hoc network consists of several nodes, depletion of batteries in these nodes will have a great influence on overall network performance. Therefore, one of the most important protocol design factors is related to device energy conservation.

To support mobile computing in ad hoc wireless networks, a mobile host must be able to communicate with other mobile hosts that may not lie within its radio transmission range. Therefore in order for one mobile host in the ad hoc network to communicate with the other not lying in its transmission range, some other hosts in its transmission range should route the packets from the source to the destination host. The conventional routing protocols used in wired networks cannot be effectively used in ad hoc networks. Hence new routing mechanisms are suggested which may be used for routing in ad hoc networks. Routing issues in ad hoc networks are beyond the scope of this chapter and are not considered here.

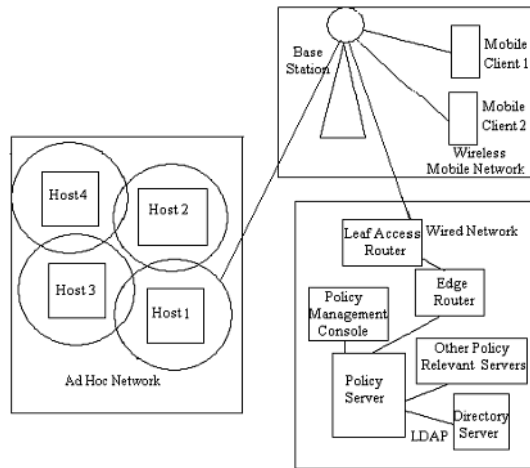
Since many mobile hosts may be within transmission range of each other, there may be multiple routes for a packet to reach a destination. Therefore the source host should decide which route to use to send the packets to reach its destination. Obviously, the sending host has to decide on the best optimal route before sending its packets towards the destination. Thus, there should be a service level agreement between the source mobile host and the host which routes the packets to the destination host. Moreover there are certain constraints based on the characteristics of the ad hoc network which play a major role in deciding which route is optimal, given there are more routes to reach the destination.

Architectural Framework of the Policy-Based Mobile Ad Hoc Network

MANET is a collection of mobile hosts forming a temporary network without the aid of any centralized administration or standard support services. The architecture for the policy-based SLAs in ad hoc networks is given below. The architecture is designed where at least one host has connectivity with the wired network. In the architecture shown in Figure 3, the policy server is placed in the wired network. Policies are stored in the directory server. The ad hoc 'host1' is within the vicinity of both 'host2' and 'host3'. 'Host4' is not within the transmission region of 'host1'. So when 'host1' wants to send a packet to 'host4', intermediary hosts, 'host2' and 'host3', help 'host1' with connection establishment.

Assuming that both the host services satisfy the constraints of 'host1', 'host1' must choose a service level agreement among the two. In this case since 'host1' is connected to a base station, which in turn is connected to the wired network having the policy server, 'host1' can query the policy server through the base station

Figure 3. Architecture of policy-based MANET



and then the leaf access router and edge router. For simplicity we have shown the policy server being connected to the base station through only a few hops. But in practice it may be many hops away from it. Once the request reaches the policy server, it takes appropriate policies from the directory server through the LDAP.

The policy server also communicates with other relevant policy servers such as Time Servers, Certificate Servers, and AAA servers, and validates the host providing service by means of certificates and AAA. The policy server makes the decision on whether the host providing the service is an authenticated one, and his services are authorized with accountability and certificates. Then the policy server based on the higher level policies stored in the directory server chooses an agreement among the available agreements. The decision to choose an agreement from among the available ones may be done giving more weight to those performance metrics which affect the overall performance the most. Over a period of time, the history of the hosts providing the service will be stored; solutions based on a neural network model may be used for finding an optimal solution. Once the policy server chooses an

agreement, it sends its reply back to 'host1', which agrees for the service with the appropriate host.

Policy Based E-Supply Chain Management Architecture

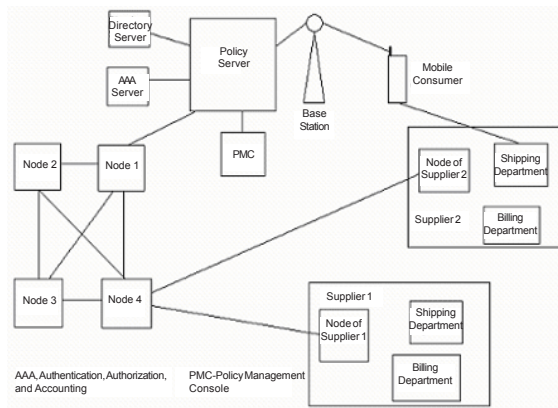
Internet-based e-purchases and e-supply chain management are now being widely used. This however has a major disadvantage of very limited mobility and the absence of a dynamic policy that will efficiently manage the entire supply chain. The activities that are required to provision mobile users comprise a surprisingly large number of steps that cross an entire enterprise. Policy setting and implementation, approval workflows, physical resource setup/teardown (provisioning), account maintenance, reconciliation of actual resource assignments with approved user lists, audit, and overall service management are some examples. They are together called policy-based provisioning. An extension to e-purchases through mobile phones using policy-based e-supply chain management is constructed.

Architecture of Policy-Based E-Supply Chain Management

The architectural framework of the e-purchase through policy based e-supply chain management is shown in Figure 4. Mobile customer, policy server, nodes (1-4), and suppliers (1 and 2) form the key elements of the proposed architecture. The mobile consumer requests the policy server of the service provider through the base station. The mobile user's authentication, authorization, and accounting rights are then verified by the AAA server. If the mobile user is found to be an authenticated one and his request for service is an authorized one, then the policy server fetches the corresponding policies for the user from the directory server through LDAP. The policy server of the ser-

vice provider is connected to other nodes (Node 1 to Node 4) in the Internet. The nodes of the suppliers such as 'supplier1' and 'supplier2' are also connected to the nodes through the public network (Internet). When the relevant policies are fetched from the directory server, the user's request is sent to an efficient supplier who will supply the product to the mobile customer. The supply chain is made electronic as discussed earlier and is policy based. Once the user's request is sent to an efficient supplier relatively nearby to the customer, the ordered products will be delivered to the mobile customer through the shipping department. The billing of the products purchased is taken into the credit account of the mobile user and is charged along with the mobile phone bill, simplifying user billing and payment. The customer at the end of the month would pay the bill through the electronic account facility available with his existing bank account. Thus the whole process of placing the purchase order, delivering the product through the supply chain, and paying for the product is made electronically, thereby facilitating the customers, who are mostly travelers and tourists.

Figure 4. Framework of e-purchase through policy-based e-supply chain management



CASE STUDY

A simulation for the policy-based MANET shown in Figure 3 was performed using the QualNet Network Simulator, using the simplex method to solve the linear program model given below, and using the Ponder Toolkit. The mathematical model is given in the following section.

Mathematical Model

A mathematical model considered for policy-based MANET uses Linear Programming and Simplex Method to solve it. The following performance metrics that are crucial for effective SLA trading and choice of route are considered in our model: (a) Bandwidth, (b) Delay, (c) Demand, (d) Packet Loss, (e) Congestion, (f) Queuing Delay, (g) Throughput, (h) Buffer Capacity, (i) Battery Consumption, and (j) Mobility. Let,

T_{ij} = Total (maximum) Bandwidth (channel capacity) available from host i to host j .

U_{ij} = Bandwidth being used for traffic flow between host i to host j at instant ' t '.

R_{ij} = Reserved bandwidth from host i to host j .

Hence the bandwidth that can be leased to other hosts G_{ij} is given by $G_{ij} = T_{ij} - U_{ij} - R_{ij}$.

Let the required bandwidth—that is, the bandwidth consumed by the host k to reach host j through host i —be RB_{ij} .

And,

D_{ij} = Delay from host i to host j .

C_{ij} = Cost of reaching host j through host i .

F_{ij} = Fraction of bandwidth bought from host i to reach host j .

The objective here is to minimize the cost of reaching host j through other hosts.

$$\text{Minimise } \sum_{i,j} F_{ij} C_{ij} \quad (1)$$

As stated earlier, in the above equation represents the cost hostcharges to reach host through host.

Constraints

There are a set of constraints that define the model. The first constraint is the demand for bandwidth to reach host j through host i ; De_{ij} should be less than or equal to the amount of bandwidth host i is ready to offer for cost to reach host j , G_{ij} .

$$DE_{ij} \leq \sum_{i,j} G_{ij} \quad (2)$$

The following constraints check if the service performance metrics in the service offered by the host i to reach host j fall within the predetermined and pre-calculated boundaries as expected by host k which needs the service. These boundary constants for the performance metrics can also be set dynamically and SLA negotiated accordingly.

Buffer Capacity B_{ij} should not be less than a bearable value given by the constant N =Number of packets that can be buffered.

$$B_{ij} \geq N \quad (3)$$

The time delay D should be set to a limit expressed by a constant ' $p1$ ' as expected by the 'ISP k ' which needs the service. The constant ' $p1$ ' is arrived as derived as follows:

' $p1$ ' = Propagation Time + Transmission Time + Queuing Delay (+ Setup Time)

Propagation Time: Time for signal to travel length of network

= Distance/Speed of light

Transmission Time = Size/Bandwidth

Therefore, we have

$$D_{ij} \leq p1 \quad (4)$$

Queuing Delay Q_{ij} should not exceed an allowable limit ' $p2$ ' expressed as

$$p2 = \frac{D}{2} \times (N - 1)$$

where, D = the time delay, N is the Buffer Capacity

$$Q_{ij} \leq p2 \quad (5)$$

The Packet Loss P_{ij} for the service provided should not exceed a maximum limit set as constant ' $p3$ ', and Congestion in the channel offered for service Co_{ij} should also be within the acceptable limits represented by the constant ' $p4$ ', both of which are arrived at as shown as follows:

T_{min} = Minimum Inter-Arrival Time observed by the receiver.

P_0 = Out of order packet.

P_i = Last in-sequence packet received before P_0 .

T_g = Time between arrival of packets P_0 and P_i .

n = Packets missing between P_i and P_0 .

If $(n + 1) T_{min} \leq T_g < (n + 2) T_{min}$, then n missing packets are lost due to transmission errors and hence ' $p3$ ' = ' n ' and

$$P_{ij} \leq p3 \quad (6)$$

Else n missing packets are assumed to be lost due to congestion and hence ' $p4$ '= n ' and

$$Co_{ij} \leq p4 \quad (7)$$

Throughput TH_{ij} should be greater than or equal to ' $p5$ ', which is given by

$$p5 = \{MSS / RTT\} \times C / (\sqrt{p})$$

where,

MSS = Maximum Segment size in bytes, typically 1460 bytes.

RTT = Round Trip Time in seconds, measured by TCP.

p = Packet loss.

C = Constant assumed to be 1.

$$TH_{ij} \geq p5 \quad (8)$$

The jitter J_{ij} should be within the acceptable limit ' $p6$ ' given by

$$p6 = p6 + (|D(i-1, i)| - p6) / 16$$

given

where, S_i, S_j are sender timestamps for packets i, j and R_i, R_j are receiver timestamps for packets i, j .

Therefore

$$J_{ij} \leq p6 \quad (9)$$

The Battery Consumption BC_{ij} for the offered service should be within the boundary constant ' $p7$ ',

$$BC_{ij} \leq p7 \quad (10)$$

The Mobility Factor M_{ij} which gives the idea of how long the host j will be in the transmission range of host i for which packets need to be routed should not be smaller than a particular constant represented by ' $p8$ ',

$$M_{ij} \geq p8 \quad (11)$$

This mobility factor M_{ij} plays a crucial role in ad hoc networks because the hosts are all mobile. It may be minutes or in any preferred time unit as the case may be. We generally assume that a mobile which has joined the ad hoc has more probability of staying in the network than the ones which came earlier than that. But the exact nature of the mobility of a host can be predicted only based on past performances of the mobile.

Non-Negativity Constraints

The following are the non-negativity constraints applied in the model:

Cost C_{ij} should always be positive,

$$C_{ij} \geq 0 \quad (12)$$

Fraction of bandwidth bought from host i to reach host j , F_{ij} should also be positive,

$$F_{ij} \geq 0 \quad (13)$$

The bandwidth that can be offered for cost to other hosts by host i should be positive,

$$G_{ij} \geq 0 \quad (14)$$

Given the objective, for example, to minimize the agreement cost along with the performance metrics constraints, the proposed linear programming model solved using simplex method suffices for arriving at a suitable agreement for service with other hosts. There are always cases that the above model will fetch more than one solution if other solutions exist. Hence in such cases the decision of choosing the most appropriate of the available solutions should be taken which is described in the next section.

The test environment has four ad hoc hosts from 'host1' to 'host4', as shown in Figure 3. The total bandwidth, used bandwidth, reserve bandwidth, battery consumption, mobility factor, and other performance metrics of the hosts are tabulated below.

Table 1. Performance metrics and other parameters of the hosts

Performance Metrics	Host 1	Host 2	Host 3
Total Bandwidth Allocated (MBps)	3	6	5
Bandwidth Used at Instant (MBps)	2	2	1
Reserve Bandwidth (MBps)	0	1	1
Remaining Bandwidth G_{ij} (MBps)	1	3	3
Demand for Bandwidth to Reach 'host4' (MBps)	1	0	0
Delay (x 10-3/sec)	7	8	10
Packet Loss Factor	7	5	6
Congestion Factor	30	20	25
Queuing Delay (x 10-4sec)	8	7	10
Throughput (x 103 Bits/sec)	100	100	90
Buffer Capacity (No. of Packets)	9	10	8
Battery Consumption (mWh)	-	8	9
Mobility Factor—Minutes	-	25	18

In the simulation test environment, 'host1' needs to communicate with 'host4', which is not in its transmission range. So both 'host2' and 'host3' offer the service to 'host1'. Using the mathematical model proposed, 'host1' decides upon the suitable service among the offers using the SLA trading algorithm (Rajeev, Sivanandam, Sreenaath, & Bharathi Manivannan, 2005) and the mathematical model given previously. Since only the service offered by 'host2' adheres to the performance metric constraints, 'host1' chooses the service offered by 'host2'. All the simulation is done with respect to the packet flow from 'host1' to 'host4'.

The trade for the service is decided by using the simplex method to solve the linear programming model and SLA trading algorithm, by which a feasible solution is obtained. The performance constraints and other parameters of the hosts are given in Tables 1 and 2. According

Table 2. Performance metrics and other parameters of 'host2' and 'host3'

Performance Metrics	Host 2	Host 3
Delay (x 10-3/sec)	2.9	3.1
Packet Loss Factor	0.2	0.3
Congestion Factor	0.3	0.3
Queuing Delay (x 10-4sec)	0.2	0.2
Throughput (x 103 Bits/sec)	4.2	4.2
Buffer Capacity (No. of Packets)	20	15
Battery Consumption (mWh)	8	9
Mobility Factor (minutes)	25	18
Jitter (x 10-4sec)	3.9	4.1
Fraction of Bandwidth that Can Be Given F_{ij} (MBps)	1	1
Cost C_{ij} (\$)	2	6

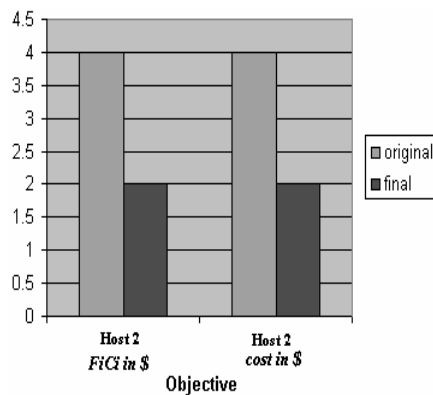
Table 3. Original and final value of the objective

Objective	Original Value	Final Value
$\sum_{i,j} F_{ij} G_{ij} C_{ij}$ of 'host2' (\$)	12	6
Cost of 'host2' (\$)	4	2

to the constraints given by 'host1' for the required service, the simplex method and SLA trading algorithm are used, and the best bid among the bids offered by the two hosts ('host2' and 'host3') is selected. Since only the bid for the service offered by 'host2' satisfies the constraints of 'host1', SLA between 'host1' and 'host2' takes place. The LP model is solved by using the simplex method. As only the trade provided by 'host2' satisfies all the constraints with the objective of minimum cost, the Service offered by 'host2' is agreed upon for trade.

From the performance metrics and the constraints on performance metrics, the objective of minimizing cost is arrived at (see Figure 5 and Table 3). Thus an effective SLA is traded between 'host1' and 'host2', satisfying the constraints on the performance metrics which affect the service.

Figure 5. Objective



CONCLUSION AND FUTURE DIRECTIONS

Policy computing can be effectively used in mobile computing in various arenas such as QoS, security, SLA, and e-purchase. The architectural framework demonstrated in the case study gives insight as to how QoS, SLA, and security can be implemented in mobile networks. Policy-based architectures for billing in mobile networks are currently being constructed which could bring transparency in mobile billing with added dynamicity.

REFERENCES

- Agrawal, P., Chen, J. C., & Sivalingam, K. M. (1999). *Energy efficient protocols for wireless networks*. Norwell, MA: Kluwer Academic Publishers.
- Braden, R., Zhang, L., Berson, S., Herzog, S., & Jamin, S. (1997). *Resource ReSerVation Protocol (RSVP)—version 1 functional specification*. IETF RFC 2205.
- Chan, K. et al. (2001). *COPS usage for policy provisioning (COPS-PR)*. IETF RFC 3084.
- Chan, K., Sahita, R., Hahn, S., & McCloghrie, K. (2003). *Differentiated Services quality of service policy information base*. IETF RFC 3317.
- Fankhauser, G., Schweikert, D., & Plattner, B. (1999). *Service level agreement trading for the Differentiated Services architecture*. Technical Report No. 59, Computer Engineering and Networks Lab, Swiss Federal Institute of Technology, Switzerland.
- Hodges, J., & Morgan, R. (2002). *Lightweight Directory Access Protocol (v3): Technical specification*. IETF RFC 3377.

Lewis, L. (1996). Implementing policy in enterprise networks. *IEEE Communications Magazine*, 34(1), 50-55.

Rajeev, S., Sivanandam, S. N., Sreenaath, K. V., & Bharathi Manivannan, A. S. (2005). Policy-based SLA for wireless ad hoc networks. In *Proceedings of the International Conference on Services Management*, India.

Sivanandam, S. N., Santosh Rao, G., Pradeep, P., & Rajeev, S. (2003). Policy-based architecture for authentication in wireless Differentiated Services using Distributed Substring Authentication Protocol (DSAP). In *Proceedings of the International Conference on Advanced Computing*, India.

Section IX

Customer

Chapter XLIV

Investigation of Consumer Behavior in Using Mobile Payment Services—A Case Study of Mobile Recreational Services in Taiwan

Maria Ruey-Yuan Lee
Shih Chien University, Taiwan

Yi-chen Lan
University of Western Sydney, Australia

Hsiang-ju Su
Shih Chien University, Taiwan

ABSTRACT

The growing popularity of the mobile phone and the diverse functionality of mobile services have forced mobile service providers to enter into a highly competitive business arena. In digital life today, mobile phone services are not restricted merely to communicating with people but more and more value-added services have emerged to amalgamate disparate industries/businesses and open up greater market opportunities. These disparate industries/businesses may include recreational and travel services, mobile learning services, mobile banking services, and many others. Nevertheless the service providers must understand the consumer behaviour in value-added services in order to enhance their product design. The key objectives of this research is to investigate and analyze the relationships between the consumer behaviour, consumer personality and lifestyle in adopting mobile recreational services; and provide recommendations to the service providers for increasing competitiveness—in the context of Taiwan.

INTRODUCTION

The rapid evolution of mobile phone technologies and services provide consumers with enormous interests in using mobile phones for many other daily activities. The service providers are racking their brains to develop increasingly value-added services to attract consumers. Today, consumers anticipate the products and services they purchased are in personalized form. Consequently mobile phone services cannot be restricted merely in communication function, and should be customized and characterized in accordance with a consumer's personality and individuality.

According to the statistic report announced by the Directorate General of Budget, Accounting and Statistics, Taiwan¹ in March 2004:

1. Total mobile phone accounts (25 million accounts) at the end of 2003 increased 5% from the previous year.
2. GPRS (General Packet Radio Service) accounts (2.68 million accounts) at the end of 2003 increased six times from the previous year.
3. There were an average of 111 accounts per 100 users at the end of 2003.
4. The total mobile phone communications duration at the end of 2003 was 23.3 billion minutes, which was a 16.6% increase from the previous year.

Based on the above figures, there is no doubt that the mobile telecommunication market is continuously growing. Furthermore, swift growth of GPRS accounts shows that the usage of mobile phones has expanded to other value-added services such as the ability to access and operate Internet applications, and the ability to remotely access and control in-house appliances and machines. For this reason, the study focuses on the following two aspects:

1. the investigation and analysis of consumer behavior of adopting value-added mobile services; and
2. the relationships with specific consumer characteristics such as personality, lifestyle, and corresponding demographic parameters (age, gender, education level, and occupation) to identify their implications.

This research concentrates on consumers adopting value-added mobile services in terms of downloading ring tones and images. Through an online questionnaire, data related to personality, lifestyle, and experiences of downloading ring tones and images can be collected and further analyzed for service providers in decision making and strategic planning.

The above research objectives are converted into the following specific research activities, which have been addressed during this study:

1. Analyze value-added mobile services users' (VAMS user) personalities, lifestyles, and their relationships with demographic parameters.
2. Evaluate the associations between VAMS users' personalities and product/service
3. Evaluate the associations between VAMS users' lifestyles and product/service categories.

Figure 1. Research framework

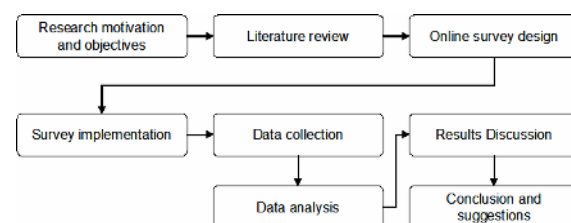


Table 1. Comparison of four generations of mobile phone technologies (based on Yu, 2002)

Generation	1G	2G	3G	4G
Main Technologies	AMPS, TACS, NMT	GSM, CDMA	UTRAN	Under development
Type of Transmission	Analogue	Digital	Digital	Combination of network and communication infrastructure
Support	Basic mobility	Roaming services	Integrated roaming services	Integrated roaming services
Bandwidth	Low, voice only	Low, voice and digital signals	Hi, digital signals and multimedia	Hi, large and real-time multimedia
Transmitting Speed	Very slow	GSM-9.6Kbps CDMA-14.4Kbps	Up to 2Mbps	Up to 100Mbps

Table 2. Comparison of characteristics between mobile business and e-business (based on Liu, 2002)

	Mobile Business	E-Business
Communication Technologies	Connecting with clients via mobile and wireless communication technologies	Connecting with clients via traditional telecommunication technologies
Scope of Services	Provides location-based information	Limited Web site-based information
Equipment for Services	Mobile phone, PDA	Desktop, laptop
Locations of Acquiring and Using the Services	Clients can use mobile business services at any time and any location	Due to fixed-point access, clients can only use the services in a single location at any one time

Furthermore a research framework is illustrated in Figure 1 to enable readers to capture the overall picture of this study.

The rest of the chapter is organized into four main sections. Firstly, the literature survey section provides a comprehensive review on the evolution of mobile phone technologies, mobile business operations, and consumer behavior on mobile services. Secondly, the research method and framework section highlights the research conceptual model, hypotheses, limitation, and questionnaire design. Thirdly is the discussion of analyzed survey results, and finally the chapter concludes with suggestions and future directions.

LITERATURE SURVEY

In order to understand and enhance mobile phone value-added services, it is imperative to review the fundamentals and evolution of mobile phone technologies, mobile business services, value-added mobile services, and mobile consumer behavior. This section has been subdivided into various corresponding sub-sections to accommodate the aforementioned components.

Fundamentals and Evolution of Mobile Phone Technologies

To give readers an overall picture of the fundamental mobile phone technologies, the authors

have summarized the comparison of four generations (1G, 2G, 3G, and 4G) of mobile phone technologies in Table 1.

Mobile Business Services

Although both mobile business and e-business services are based on the Internet platform, there are still quite a few dissimilarities between them. Table 2 summarizes the differences between mobile business and e-business in terms of their communication technologies, scope of services, equipment for services, and locations of acquiring and using the services.

Mobile business services may include the following features:

1. Users can access the services regardless of time and location limitation—that is, “ubiquity.”
2. Easy and convenient—users are not required to go through a sophisticated procedure to complete transactions.
3. Mobile business services can provide users with real-time information, such as such as traffic report, weather conditions, and locations of ATMs, based on the users’ location.
4. Mobile business services allow users to select functions based on their personal preferences and setup the services details—that is, “personalization.”
5. Users can access the services network (e.g., mobile banking) without complicated procedures.

Value-Added Mobile Services

At present, the mobile service providers deliver value-added mobile services in two main categories. One is aimed at end users for consuming services, another is targeted at enterprise users for systems integration services. In general, value-added mobile services can be

Table 3. Value-added mobile service categories

Service Categories	End Users	Enterprise Users
Information Services	• Reference information (e.g., maps) • Real-time updated information • Location-based information	• Data collection • Data monitoring • Alerting and reminding mechanisms
Communication Services	• Text messaging services • Multimedia messaging services	• Calendar • Messaging services • E-mail services • Group messaging
Mobile Business Services	• Retailer services • Finance and banking services • Transaction and payment mechanisms	• Customer relationship management • Sales and marketing services • Customer services
Entertainment Services	• Games, images, and ring tones’ downloading • Image transmissions	

grouped into four main categories, including information services, communication services, mobile business services, and entertainment services (see Table 3).

Mobile Consumer Behavior

According to Pearson Education’s online glossary (2000), “consumer behavior” refers to the process by which people determine whether, what, when, where, how, from whom, and how often to purchase goods and services. In other words, it is to investigate why people purchase commodities. Once sales and marketing staff recognized the reasons people purchased particular brands or products, enterprises could rely on such information to fine-tune their business strategies to enhance customer services and satisfaction.

In general, consumers make purchasing decisions through six phases. Nevertheless consumers may withdraw purchase decisions at any point. The following six phases delineate

the pattern of Internet consumers' decision procedures (Lin et al., 2002).

Phase 1—Requirement Confirmation

This phase initializes the entire purchase decision process. The purchase requirement is confirmed when consumers think, feel, see, and understand how the featured products or services will help make their life or job easier.

Phase 2—Information Gathering

There are two types of information gathering. The first type is “internal search”—consumers search their past experiences. The second approach is “external search”—consumers search for related information via external channels to assist their purchase decision making. In the traditional purchase environment, gathering information requires time and effort; however, in the Internet environment, information is searched and gathered through powerful search engine facilities and various information providers and agents. The time and effort required for Internet-based information gathering are far less than the traditional approach.

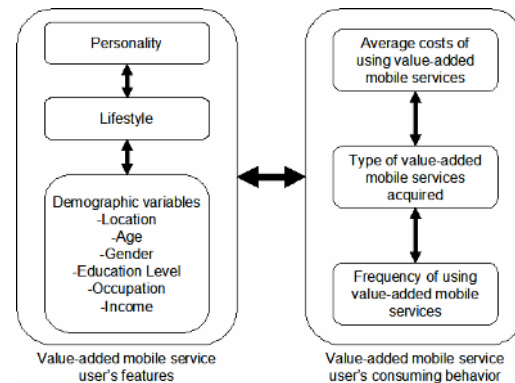
Phase 3—Evaluation of Acquired Information

After gathering sufficient information, consumers will evaluate the information to identify the appropriate solution (products or services).

Phase 4—Purchasing

At this point consumers may or may not proceed with their purchases. It is important that online merchants should recognize the unsuccessful purchase factors.

Figure 2. Conceptual model



Phase 5—After-Purchase Evaluation

After the successful purchase, consumers would compare the purchase with their expectations. After-sales services are an important evaluation element and a key factor to determining the returning consumer.

Phase 6—Returning and Repurchasing

If the after-purchase evaluation does not meet consumers' satisfaction, the consumers may return their purchase and restart their repurchasing activities.

RESEARCH METHOD AND CONCEPTUAL MODEL

After the review of mobile phone technologies, mobile business and value-added services, and consumer behavior, this section identifies an appropriate research method and develops a conceptual model (see Figure 2) to carry out the study.

Hypotheses

The following hypotheses are derived in accordance with the literature review, research framework, and research objectives:

- H1: There is a significant correlation between the VAMS user's personality and lifestyle.
- H2: There is a significant correlation between the VAMS user changing ring tones frequency and demographic variables.
- H3: There is a significant correlation between the VAMS user changing display images frequency and demographic variables.
- H4: There is a significant correlation between the VAMS user's average spending on value-added services and demographic variables.
- H5: There is no significant correlation between the VAMS user's personality and the costs of using value-added services.
- H6: There is a significant correlation between the VAMS user's lifestyle and the type of acquired value-added services.
- H7: There is no significant correlation between the VAMS user's lifestyle and changing display images frequency.
- H8: There is no significant correlation between the VAMS user's lifestyle and changing ring tones frequency.
- H9: There is no significant correlation between the VAMS user's lifestyle and the costs of using value-added services.

Research Design

An online survey was designed to allow any Internet users to answer the survey questionnaires. The survey was written in ASP with SQL backend database system and is located at <http://home1.usc.edu.tw/M9196012/main.asp>. The survey system has the capability of avoiding repeated participants and uses entered e-mail addresses to remove suspicious samples.

Constructing Questionnaires

The survey is divided into three main sections. The first section contains 35 questions relating to the VAMS user's personality. This section is based on Cohen's (1967) CAD personality measurement. The second section adopts from SRI Company's VALS2 survey (2005). It measures people's lifestyles of managing time and money. The section contains 33 questions. The third section is aimed at collecting demographic data (age, gender, education level, occupation, and income) of the VAMS user, and it contains eight questions. The survey can be found in the Appendix.

RESULTS ANALYSIS AND DISCUSSION

The duration of survey data collection was four weeks after it was launched on the Internet. A total of 138 samples derived after a validation process. A sample of 138 respondents yields a sampling error of just under $\pm 10\%$ at the 95% confidence level. Therefore, for example, if half the respondents in the sample agree with an item and half disagree, then in 19 out of 20 cases the true percentage agreeing in the wider population would range between 40% and 60% (i.e., $\pm 10\%$). That would be the maximum variance potentially arising due to sampling effects.

In addition to basic (descriptive) statistics analysis, the study also embraces Reliability Analysis (Cronbach's Coefficient Alpha), Factor Analysis, and Bivariate Correlation Analysis.

A summary of data analysis results is divided into two main groups and presented as follows.

Mobile Phone User Profile in Taiwan

1. *Age Distribution*—61.6% of the respondents are in the age group of 23-30.
2. *Gender Distribution*—There is not much difference between male (44.9%) and female (55.1%) respondents.
3. *Education Level*—89.9% of the respondents hold college degrees or higher.
4. *Monthly Allowance*—The majority of the respondents (55.8%) have a monthly allowance in the following two groups:
 - a. NT\$1,001-5,000—26.8%
 - b. NT\$5,001-10,000—29%
5. *Occupation Distribution of the VAMS User*—Student is the largest group (30.4%), followed by IT industry (13.0%). The communication industry also holds 11.6% of the VAMS users.
6. Nearly two-thirds of value-added mobile services users (69.6%) live in the northern part of Taiwan (close to the capital city of Taipei).
7. 43.5% of VAMS users have used mobile phone services for more than five years.
8. A total of 96.4% of mobile phone accounts are monthly subscription and pre-paid services.
9. The primary concern of adopting value-added mobile services of the mobile phone users is the “cost” of the services (31.9%).
10. About 42.0% of the mobile phone users learned the value-added mobile services from TV advertisements.
2. A significant negative correlation between the VAMS user’s personality and the frequency of changing mobile phone ring tones.
3. A negative correlation (not significant) between the VAMS user’s personality and the frequency of changing mobile phone display images.
4. A positive correlation between the VAMS user’s personality and the average costs of subscribing value-added mobile services.
5. A positive correlation between the VAMS user’s lifestyle and the types of subscribed value-added mobile services.
6. A positive correlation between the VAMS user’s lifestyle and considering adopting the types of value-added mobile services.
7. No correlation between the VAMS user’s lifestyle and the frequency of changing mobile phone display images.
8. No correlation between the VAMS user’s lifestyle and the frequency of changing the mobile phone ring tones.
9. A positive correlation between the VAMS user’s lifestyle and the average costs of subscribing value-added mobile services.

Overall, the characteristics of the VAMS user such as age, gender, education level, personality, and lifestyle have positive influence on the user’s consuming behavior.

Discussion

Derived from the above analysis, we identify and highlight the following recommendations for sales and marketing strategy of mobile services providers.

From the demographic variable perspective, balanced gender distribution illustrates that

- the sales and marketing campaign should not focus on either the female or male domain;

Relationships between Personality, Lifestyle, and Mobile Phone Services Consuming Behavior

1. There are significant correlations between the consuming behavior and the VAMS user’s age, education level, occupation, and the available monthly allowance.

- age distribution clearly indicates the younger generation (23-30 age group) makes up the majority of mobile phone users, hence the types of value-added mobile services should be designed to suit these age group's needs; and
- costs of adopting value-added mobile services is another major hurdle to attract new customers, consequently subscription fees of popular value-added services should be positioned within the mobile phone user's monthly allowance range.

The VAMS users' lifestyles are influenced by their personalities, and the differences of lifestyles will influence their decisions in choosing value-added mobile services. Therefore, the mobile service providers need to identify the characteristics of each kind of lifestyle, and design and package the value-added services to satisfy the requirements.

Furthermore, the results evident indicate that students are the prominent mobile phone user group. However, many service providers possess a misperception of consuming capability of the student group and fail to benefit from such a great market. The authors would like to suggest that service providers investigate the trends and needs among the student domain and design suitable value-added mobile services to attain such a market.

CONCLUSION AND FUTURE DIRECTIONS

This study aims at investigating the VAMS users' consuming behavior and identifying the correlations between the consuming behavior and various characteristics such as personality, lifestyle, and a number of demographic variables. The research findings indicate that there are significant correlations between the VAMS users' consuming behavior and various factors.

These factors have imperative impact on the decisions of VAMS users adopting and choosing value-added mobile services—namely, personal characteristics (personality) and lifestyle. A number of demographic variables are embedded into the research design to further analyze the implications of the factors from the demographic perspective. From the results discussion, the authors have recommended that mobile phone providers explore the experiences and issues faced by the existing customers, and revisit their sales and marketing campaigns and strategies.

The investigation of VAMS users' consuming behavior has led naturally to many ideas that could be further pursued by the authors, as well as other researchers who might be interested in these ideas and who embrace a similar research paradigm. Furthermore, as mentioned in the introduction, the value-added mobile services are continuously changing and evolving. Therefore, mobile phone service providers are likely to face many other issues and challenges. The following points suggest future research topics that are emerging from this study:

1. Various environmental factors may be included in future study to represent more accurate and realistic situations.
2. Value-added mobile services embrace a much wider scope than the current study. Many other services such as mobile Internet access, mobile shopping, small amount payment, and real-time and on-demand multimedia services should be considered and integrated for further analysis.

REFERENCES

Cohen, B. J. (1967, August). An interpersonal orientation to the study of consumer behavior. *Journal of Marketing Research*, 4, 270-278.

Frolick, M. N., & Chen, L. (2004). Assessing m-commerce opportunities. *Journal of Information Systems Management*, 21(2), 53-61.

Lin, et al. (2002). *Internet marketing*. Taipei, Taiwan: Flag Publishing.

MarketIntelligenceCenter. (2005). Retrieved March 3, 2005, from <http://mic.iii.org.tw/intelligence/>

Patrick, E., & William, A. (1979). A modernized family lifecycle. *Journal of Consumer Research*, 6(1), 16.

Pearson Education. (2000). *Online glossary*. Retrieved April 29, 2005, from http://www.prenhall.com/rm_student/html/glossary/c_gloss.html

SRI Consulting Business Intelligence. (2005). *The VALS survey*. Retrieved March 30, 2005, from <http://www.sric-bi.com/VALS/presurvey.shtml>

Stafford, T. F., & Gillenson, M. L. (2003). Mobile commerce: What it is and what is could be. *Communications of the ACM*, 46(12), 33-34.

Yu, F. (2002). *Mobile telecommunications*. Taipei, Taiwan: Kings Information Co.

ENDNOTE

¹ <http://eng.dgbas.gov.tw/mp.asp?mp=2>

APPENDIX

1. Have you ever used any value-added mobile services (e.g., SMS, downloading music, ring tones, or images)?
Yes (please continue) No (please stop, thank you)
2. Have you answered this survey previously?
Yes (please stop, thank you) No (please continue)

Part 1: About your personality

Please choose the answers that most suit you.

No.	Question	1	2	3	4
1	Being free of emotional ties with others				
2	Giving comfort to those in need of friends				
3	The knowledge that most people would be fond of me at all times				
4	To refuse to give in to others in an argument				
5	Enjoying a good movie by myself				
6	For me to pay little attention to what others think of me				
7	For me to be able to own an item before most of my friends are able to buy it				
8	Knowing that others are somewhat envious of me				
9	To feel that I like everyone I know				
10	To be able to work hard while others are elsewhere having fun				
11	Using pull to get ahead				
12	For me to have enough money or power to impress self-styled "big shots"				
13	Basing my life on duty to others				
14	To work under tension				
15	If I could live all alone in a cabin in the woods or mountains				
16	Punishing those who insult my honor				
17	To give aid to the poor and underprivileged				
18	Standing in the way of people who are too sure of themselves				
19	Being free of social obligations				

Investigation of Consumer Behavior

20	To have something good to say about everybody				
21	Telling a waiter when you have received inferior food				
22	Planning to get along without others				
23	To be able to spot and exploit weakness in others				
24	A strong desire to surpass others' achievements				
25	Sharing my personal feelings with others				
26	To have the ability to blame others for their mistakes				
27	For me to avoid situations where others can influence me				
28	Wanting to repay others' thoughtless actions with friendship				
29	Having to compete with others for various rewards				
30	I knew that others paid very little attention to my affairs				
31	To defend my rights by force				
32	Putting myself out to be considerate of others' feelings				
33	Correcting people who express an ignorant belief				
34	For me to work alone				
35	To be fair to people who do things which I consider wrong				

Part 2: Your lifestyle

No.	Question	1	2	3	4
1	I am often interested in theories.				
2	I like outrageous people and things.				
3	I like a lot of variety in my life.				
4	I love to make things I can use everyday.				
5	I follow the latest trends and fashions.				
6	I like being in charge of a group.				
7	I like to learn about art, culture, and history.				
8	I often crave excitement.				
9	I am really interested only in a few things.				
10	I would rather make something than buy it.				
11	I dress more fashionably than most people.				
12	I have more ability than most people.				
13	I consider myself an intellectual.				
14	I must admit that I like to show off.				
15	I like trying new things.				
16	I am very interested in how mechanical things, such as engines, work.				
17	I like to dress in the latest fashions.				
18	There is too much sex on television today.				
19	I like to lead others.				
20	I would like to spend a year or more in a foreign country.				
21	I like a lot of excitement in my life.				
22	I must admit that my interests are somewhat narrow and limited.				
23	I like making things of wood, metal, or other such material.				
24	I want to be considered fashionable.				
25	A woman's life is fulfilled only if she can provide a happy home for her family.				
26	I like the challenge of doing something I have never done before.				
27	I like to learn about things even if they may never be of any use to me.				
28	I like to make things with my hands.				
29	I am always looking for a thrill.				
30	I like doing things that are new and different.				
31	I like to look through hardware or automotive stores.				
32	I would like to understand more about how the universe works.				
33	I like my life to be pretty much the same from week to week.				

Part 3: Mobile phone usage

1. How long have you been using mobile phone services?
 - a. < 6 months
 - b. 6-12 months
 - c. 1-2 years
 - d. 2-3 years
 - e. 3-4 years
 - f. 4-5 years
 - g. > 5 years
2. What type of mobile phone account do you subscribe to?
 - a. Monthly subscription
 - b. Pre-paid
3. What is your main concern when you consider adopting images, ring tones, and call waiting tones download services?
 - a. The contents are new and fashion
 - b. The contents represent my personal style
 - c. Costs of the services
 - d. Easy to set up
4. Where did you [hear about] the value-added mobile services?
 - a. TV advertisements
 - b. Magazine and newspaper
 - c. Friends
 - d. SMS advertisements
 - e. Internet
 - f. Other, please specify _____
5. On average, how often do you change your mobile phone display image (or wallpaper)?
 - a. More than twice a week
 - b. Once a week
 - c. Once every fortnight
 - d. Once a month
 - e. Twice a year
 - f. Once a year
 - g. Never change
6. On average, how often do you change your mobile phone ring tones (or call waiting ring tones)?
 - a. More than twice a week
 - b. Once a week
 - c. Once every fortnight
 - d. Once a month
 - e. Twice a year
 - f. Once a year
 - g. Never change
7. How much do you pay for value-added mobile services every month?
 - a. < NT\$100

Investigation of Consumer Behavior

- b. NT\$100-NT\$200
 - c. NT\$200-NT\$300
 - d. NT\$300-NT\$400
 - e. NT\$400-NT\$500
 - f. NT\$500-NT\$1,000
 - g. > NT\$1,000
8. Who is your mobile phone services provider?
- a. China Telecom
 - b. FETNet
 - c. Taiwan Mobile
 - d. KGT
 - e. MOBILAI
 - f. TransAsia
 - g. PHS
 - h. APBW
 - i. Other, please specify _____

Part 4: Personal details

1. Gender
- a. Male
 - b. Female
2. Age
- a. < 12
 - b. 13-18
 - c. 19-22
 - d. 23-25
 - e. 26-30
 - f. 31-35
 - g. 36-40
 - h. 41-50
 - i. 51-60
 - j. > 60
3. Education Level
- a. Primary School
 - b. Junior High School
 - c. Senior High School
 - d. College and University
 - e. Master and above
4. Occupation
- a. Student
 - b. IT Industry
 - c. Manufacturing (exclude IT Industry)
 - d. Teacher/Professor

- e. Public Sector
 - f. Business and Trade
 - g. Finance and Real Estate
 - h. Public Broadcasting/Advertisement
 - i. Building
 - j. Telecommunication
 - k. House Administration
 - l. Retired
 - m. Other, please specify _____
5. Monthly Allowance (available for any purposes)
- a. < NT\$1,000
 - b. NT\$1,001-NT\$5,000
 - c. NT\$5,001-NT\$10,000
 - d. NT\$10,001-NT\$15,000
 - e. NT\$15,001-NT\$20,000
 - f. NT\$20,001-NT\$50,000
 - g. > NT\$50,001
6. Where do you live (which part of Taiwan)?
- a. North
 - b. Centre
 - c. South
 - d. East
 - e. Islands

Chapter XLV

Mobile CRM: Reaching, Acquiring, and Retaining Mobility Consumers

Chean Lee

Methodscience.com, Australia

ABSTRACT

This chapter provides an introduction of using Mobile CRM to reach, acquire, convert and retain consumers. Firstly, a definition of the term CRM is provided and the author also gives an insight on extending CRM to the wireless world. Having presented the benefits of mobile data services and their benefits to businesses in terms of customer relations and marketing, however, businesses still faced the challenges on delivering the promise to consumers. More importantly, the adoption of mobile services is still low in business and consumer segments. The author identifies content appropriateness, usability issues, personalization, willingness to pay, security and privacy as major challenges for businesses, and then, recommends businesses to start segmenting their mobile consumers into: Mobile Tweens, Mobile Yuppro and Senior Mobile users and the understanding of demographics, social and behavioural issues of these three consumer groups as initial step in Mobile CRM, before finally recommending the use of viral marketing as a mechanism to market mobile services. This is followed by matching relevant services to consumers create positive usability experience and always build a critical mass but develop a customer at one time.

INTRODUCTION

In recent years, we have seen an explosion in mobile entertainment, mobile B2B applications, mobile devices, and the wireless Web access packages from the telcos. This chapter discusses the utilization of CRM (customer rela-

tionship management) in reaching, acquiring, converting, and retaining mobile consumers. This chapter, therefore, includes discussions on mobile CRM strategy, market segmentation, and applying mobile CRM in the customer engagement process.

The term customer relationship management encompasses many descriptions which depend on different purposes. In the early days of CRM, business application providers often related CRM to sets of technology modules which included: marketing automation (i.e., campaign management, Web analytics, market forecasting); sales force automation (i.e., opportunity management, quotation generation, sales analytics, etc.); call centre application; order management system; and partner relationship management (PRM) that automates the entire customer management cycle.

The CRM is a business strategy that turns customer data into insights and provides a profitable process in handling client relationships. CRM technology packages act as tools to enhance the entire business strategy and processes.

The core objective of this chapter is to extend the CRM approach to mobility and the Wireless Web. In short, mobile CRM could consist of the following core components:

- Short Message Service-based advertising;
- mobile opt-in for customer data acquisition;
- mobile coupon or redeemable m-voucher;
- personalized mobile portal that offers content ranging from ring tones, news, horoscope, m-payment, and others;
- mobile alert function; and
- Web-based mobile campaign management engine.

THE POSSIBILITY OF MOBILE CRM

With the abundant availability of mobile commerce and technology, the questions arise on the benefits of mobile application in shaping the new customer relationship experience. Ac-

cording to research conducted by the management consulting firm McKinsey & Company in Europe, SMS is an effective mechanism to boost ratings and advertising sales to TV broadcasting. In the study, McKinsey pointed out that by linking TV broadcasting with SMS platform to certain shows, it enables cable TV broadcasters to boost audience loyalty. In some cases, the addition of SMS boosted the viewership of popular free-to-air television shows by up to 20% (Bughin, 2004).

In addition, SMS has also become a chatting medium for audiences ranging from ages 16-30. With the evolution of mobile technology, multimedia messaging services (MMSs), Java Games, mobile portals, and so forth, many opportunities have been created for carriers, marketers, telecoms, and businesses in building customer loyalty. Last but not least, SMS marketing from carrier and portal provider has become an acceptable advertising format compared to magazines, direct mail, or telesales, according to the mobile marketing solutions provider Enpocket (2005).

Despite the bright future of mobile technology, the adoption of mobile CRM as a form of relationship building tool is still facing obstacles. We will present some challenges faced by carriers, content providers, middleware companies and mobile marketing agencies, or even businesses in implementing mobile CRM.

CHALLENGES OF GENERATING DEMAND AND BUILDING RELATIONSHIPS WITH MOBILITY CUSTOMERS

It is reported that over \$180 billion has been spent on 3G licenses for developing the next generation of mobile data services, which is believed to have attracted 10 million regular users in the UK alone, according to research

conducted by the management and IT consulting firm, Accenture (2001a).

These services offered by the mobile service provider seem hard to impress and capture the attention of mobility consumers. So what could be classified as the potential obstacles and challenges in offering a range of wireless Web and data solutions to customers?

There are several factors that caused mobility customers to dropout from a range of wireless services:

- content appropriateness and freshness;
- usability issues;
- personalization;
- willingness to pay; and
- privacy, spam, and security.

Content Appropriateness and Freshness

Appropriateness refers to the relevancy of mobile content and data services to the consumers. Freshness means the frequency of mobile content updates. For example, if a person is attracted to horoscopes, he/she must receive alerts of career, fortune, and luck of the Scorpio to his/her mobile device every morning before going to work. The quality and the daily update of the content are very important for these individuals to continue their subscription to horoscope mobile content.

Usability

Mobile application or services are often limited by screen size, and mobile users are often not very technical people to operate the PC and Internet browser in a perfect manner. As a result, simplicity in design limits the layer of mobile phone keypads operations, the completion of a WAP page navigation cycle, or even

the reduction of the number of scrolls to complete an air ticket reservation using SMS.

Personalization

Can I create my own mobile “universe” in a mobile portal? This refers to personalization whereby a mobility consumer creates his/her favourite contents like news alert, stock market updates, ESPN sports on the mobile, and others.

Willingness to Pay

Questions often arise whether consumers are willing to continuously subscribe to mobile content such as games, songs downloads, and other related leisure and entertainment. This is often a challenge to the conversion and retains the cycle of a mobile content provider.

Privacy, Spam, and Security

Do mobile marketers, ad agencies, mobile network operators, and content providers respect the privacy in obtaining opt-in data from consumers for a range of mobile services? Nowadays, consumers are bombarded by junk mail and SMS promotions; these will lead to drop-outs of consumers for mobile marketing messages. In addition, security is an important issue to be solved in offering m-payment and m-banking.

In the next section, we provide a framework for mobile service providers in building the relationship in a wireless world. Practitioners are encouraged to customize the framework that suits their business objectives, marketing, and CRM strategy. Because the write-ups in this chapter serve as a blueprint in building a mobility relationship, they need to be adjusted and amended when dealing with different scenarios.

Table 1. Mobility consumers segments

User Groups	Demographic	Character	Lifestyle	Mobility Motif
Mobile Teens	Aged between 9-20	Tech-savvy innovators who are willing to try on new mobile gadgets and services.	Enjoy variety in entertainments, live and breathe on own communities, love social and network games.	Entertainment and online community participation.
Mobile Yuppro	Aged between 26-40	Rational users used the Internet and mobile frequently in improving job productivity as well as entertainment.	Work hard and play hard. Seeking balances between leisure and career.	As a productivity tool for office administration, work scheduling, save time, and also entertainment.
Senior Mobitizen	Aged 50+	Realistic, used mobile and Internet for researching finance package, welfare, and e-mail. Cautious on trying new technologies.	Have time to spare, enjoy quality lifestyle after retirement.	Mobile is a tool for me to receive target ads on finance, best bargains, etc. for improving my life and family lifestyle.

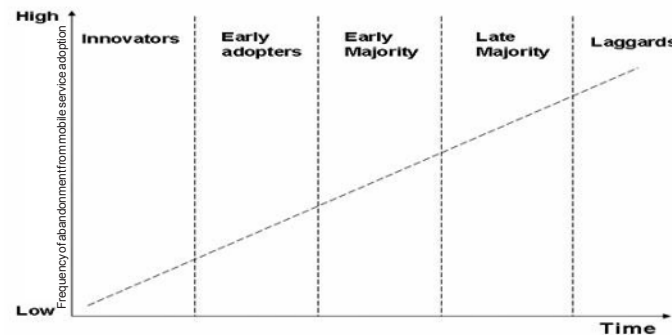
MOBILITY CUSTOMER SEGMENTATION AND POSITIONING

Before implementing mobile CRM to acquire and build customer relationships in the wireless world, it is important for us to look at the customer segments and mobility motives. Based on findings from various researches (typically, Nokia and Sony Ericsson), we can divide mobility consumers into three major groups—*Mobile Teens* (teenagers and youngsters who used mobile phones frequently), *Mobile Yuppro* (young, urban professionals that used Mobile Internet), and *Senior Mobitizen* (senior consumer groups). Table 1 provides a summary of the characteristics, usage, motives, and sets of mobile CRM strategy and applications to be offered. The summary is based on the Scenario Grid by Lindgren, Jedbratt, and Svensson (2002).

Figure 1 segments the mobility consumer group we identified in Table 1 into five major adoption phases:

- **Innovators:** This segment refers to consumers described as techies. They are pioneers in technology adoption throughout the marketing cycle. They are influencers that help mobile application, content, and service providers in developing new products.
- **Early Adopters:** This segment refers to consumers willing to try out new options, as well as visionaries that believe mobile services will boost their work productivity as well as entertainment lifestyle.
- **Early Majority:** This segment refers to consumers who possessed a “wait and see” buying behaviour. They are sceptical of mobile services and applications which really create benefits to them. They often conduct research and ask for referral before the adoption of new services.

Figure 1. Phases of mobile services adoptions (based on Accenture, 2001a)



- **Late Majority:** This segment refers to consumers that are conservative towards new mobile technologies.
- **Laggards:** This segment refers to consumers that are hard to change and resist changing. It takes time to convert them into new mobile technology adopters.

From this analysis, we can conclude that mobile services adoption is seen as an interactive relationship between suppliers and consumers. Thus, there is urgency for service providers to provide the relevant product to the users. It is important to understand the consumers' need in the wireless world, and we are now entering a "pull marketing" era. The following are CRM strategies that mobile marketers and service providers adhere to in order to reach, acquire, and retain targeted audiences:

1. **Understand Customers and Make Use of Viral Marketing:** Get to know your customers often affected by the overall CRM campaign used to communicate and interact with them. First, understand the overall needs of consumers in relation to their demographics, age, lifestyle, and adoption behaviour as mentioned in Table 1; Figure 1 helps organizations implement a successful mobile CRM in acquiring customers. Secondly, do not ignore the

power of viral marketing, which is the creation of positive word of mouth or chain reaction in mobile eco systems. This means that we need to build a critical mass of product innovators, early adopters, or even early majority user groups, as we defined in Figure 1. This is due to the fact that innovators, early adopters, and the majority often are opinion leaders; they often act as a reference for the late majority or even conservative customers. Figure 2 illustrates the power of viral marketing plays as a mobile CRM tactic.

2. **Product Matching:** Offering relevant mobility products and services that matched the needs of the consumers.

Figure 2. Viral marketing in mobile ecosystem (based on Accenture, 2001a)

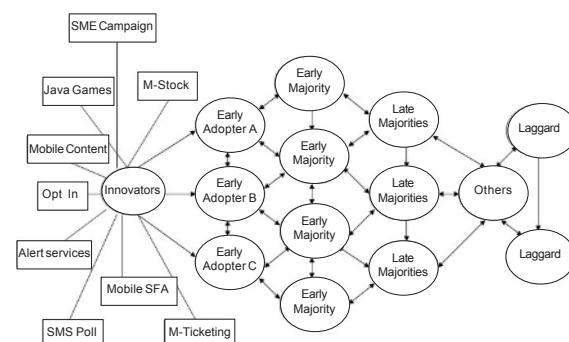


Table 2. Example of mobile services offering based on behavioural mode

Behavioral Mode	Characteristics	Example of Services	Mobile CRM Methods for Reaching Consumers
Productivity	Increase work productivity and save time.	Mobile personal organizer, e-mail, mobile sales force automation, etc.	Mobile Portal, mobile opt-in application for gathering client information regarding their interest; permission-based marketing, campaign.
Social and Entertainment	Entertainment and socialize with friends.	Mobile Java games, ring tones.	Mobile game engines for tracking the number of game downloads, winning points tracking, viral campaign.
Mobility	Get information at any time and anywhere.	Locations and positioning services.	Alert services subscription. Example, notification on bank account balance via SMS

Before mobile service providers offer more personalized products to consumers, they can always start with basic features like bulk SMS, mobile e-mails, personal organizers, and others. In the second step, we can further divide our segments we identified in Table 1 in terms of “behaviourgraphic” of mobility consumers. Table 2 illustrates the possible services as well as mobile CRM methods in reaching consumers with different behaviourgraphics.

We have seen the possibility of CRM tactics in reaching, acquiring, and retaining customers based on their behaviour. Remember, it is recommended that network operators, content developers, technology companies, or even marketing agencies should provide an open platform that enables partners to offer their services to consumers—for example, the Nokia developer networks, Telenor Mobile open platform and NTT Do Como open platform.

3. **Create User Experience:** When I was an e-business analyst at a Web development company, I always spoke to my clients about usability design and positive user experience on Web sites and Web-based applications. With the emergence of Mobile Internet and application as a

mechanism in improving work productivity, entertainment options, and m-commerce, it is important for marketers to design their mobile services in usability standard. Creating positive user experience is an important CRM strategy in mobility space. Think about the available characteristics of connection protocols like WAP, GPRS, EDGE, and UMTS: screen size issues; differences on mobile gadgets like smart phones, Pocket PCs, Palm Pilots, and mobile notebooks. As a result, design with standards is crucial to create a positive experience. User experience is about accessible, content freshness, relevance, and personalization.

4. **Develop One Customer at a Time:** In the mobile universe, reaching consumers with SMS campaigns and mobile coupons, followed by relevant content to acquire consumers, is not the end of the journey. The key to mobile success is retaining the consumers. The following are the CRM strategies that help to develop one-to-one marketing for mobility consumers:

- Understand the consumers by segmenting the users based on demographics, behaviours, lifestyle, and motivations.
- Practice permission-based marketing to gather consumer information

on preferences about mobile services. Respect privacy and do not create mobile spam.

- Offer customer education and relationship building with relevant content and personalization.
- Retain the consumer by influencing and changing their spending behaviour. Build a mobility Internet community to create a mobile value chain.

Now that we have a clear blueprint and framework about pull marketing for our mobile CRM strategy to reach, acquire, and retain consumers, we will provide an in-depth walkthrough of the consumer engagement process.

THE CUSTOMER ENGAGEMENT PROCESS

The focus on the customer engagement process in the mobile world and understanding of the process will enable us to apply the pull marketing concept to each phase of the process. We can also further analyse the dropout risks in the process and provide examples of applying mobile CRM in each phase with the provision of user scenario or persona. Measurement metrics on the effectiveness of mobile CRM will be provided.

So how do we define the customer engagement process? It refers to the series of stages where a customer converts to a purchase of a product and services offered by an organization. It consists of four phases: Reach, Acquire, Convert, and Retain. Figure 3 shows these phases, modified based on the online consumer engagement process by Sterne (2002).

Reach

Reach in a mobile world refers to an informative stage where wireless marketers, mobile network operators, content providers, software companies, and mobile manufacturers utilized communication mediums to provide information for consumers—for example, an SMS marketing campaign to alert customers to participate in mobile game contests to win free movie tickets. In a mobile world, reach also must be built based on customer permission.

Acquire

This stage involves the effectiveness of the outcomes of the first stage, where mobility consumers are alerted by the SMS marketing campaign. Further interest is built between marketers and consumers.

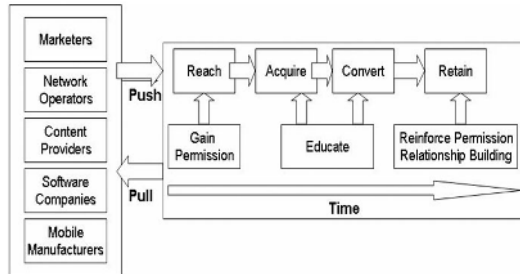
Convert

This refers to the persuasion stage of the SMS campaign or even MMS content. It involves the participation of users in response to the SMS campaign. For example, the download of mobile Java games, the use of mobile a phone to book an airline ticket, or even the purchase of movie tickets using SMS.

Retain

When the mobile consumers have the experience and familiarity with the applications, they will become frequent users of the application. For example, a consumer will continue to use the SMS as a tool for movie ticket purchasing. In this stage, a vital factor is where the satisfied customer will spread positive messages to their friends regarding the mobile application or cam-

Figure 3. Customer engagement process in mobile world



paign. It is also important for marketers to reinforce the permission and retain customers by offering fresh and relevant mobile content and services.

We conclude from Figure 3 that in wireless world, reaching and acquiring customers utilizing CRM methods like SMS or MMS ad messages and content are not the destiny; relationship building is the key to success in the customer engagement cycle. Let us apply the famous 80/20 rule into the wireless world. We can say 80% of the mobile content provider sales revenue is generated from 20% of the customers. Thus, we can see how important ongoing customer retention is in the mobile world. Again, customer acquisition is about reaching mass audiences, but developing selected target consumer groups over time. When I was involved in Pre Sales in an ERP solutions company, it took much cold calling from my telesales colleague to generate leads, and then I had to determine that the leads were strong during my client engagement. From there I prepared proposals to turn them into my prospects, and I offered demos to turn them into hot prospects. I closed a few of them from the hot prospect lists. So we can see the cycle is like a funnel: it involves some key metrics and CRM strategies we apply in each stage, is like playing with numbers, and builds up targeted long-term customers (see Figure 4).

Figure 4. Mobility customers conversion funnel

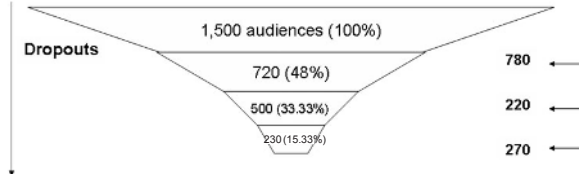
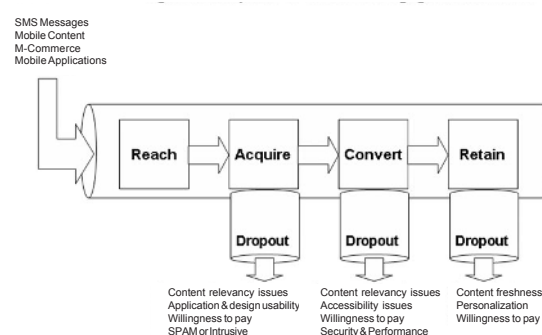


Figure 4 illustrates an example of a mobile marketing campaign. During the process, the marketer successfully sends 1,500 SMS marketing messages to 1,500 audiences, and during the entire consumer engagement cycle, we expect leakage and dropout of consumers during each stage, which could be due to irrelevant mobile content, accessible issues, negative user experience, and so on. In the final stage, we successfully converted 230 out of 1,500 users into our customers. Thus, we can see that conversion in a wireless world is also a numbers game; however, it is important for marketers to practice the concept of relationship building in the CRM context.

DROPOUT RISKS OF CUSTOMER ENGAGEMENT CYCLE

We have already discussed the issues that cause the dropout of consumers from wireless

Figure 5. Dropouts in customer engagement process



services. Figure 5 shows the various leakage or dropout risks that happen in the engagement process. This figure is modified based on the customer retention process addressed by Sterne (2002).

The challenges and factors for the adoption of mobile services are already identified. Figure 5 shows how to apply these factors to each phase of the customer engagement process. This figure summarizes that overall dropout in several phases are due to content relevance and freshness, application design usability, willingness to pay for mobile services, and security and privacy issues.

In the next section, we will provide a real-life user scenario, best industry practices, and measurement metrics for each phase in the customer engagement process.

APPLYING THE FRAMEWORK: DEFINED MARKETING OBJECTIVES AND CRM CAMPAIGN DESIGN

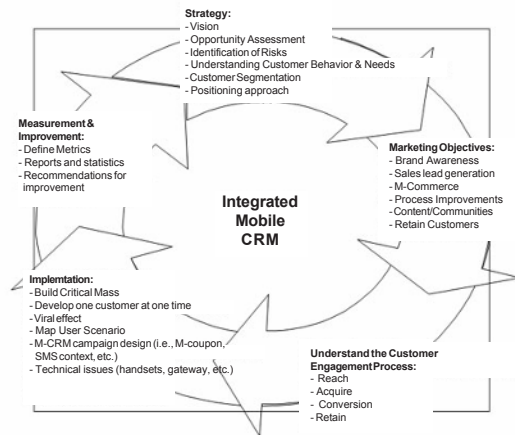
It is important to have a clear objective when we carry out a specific task or project. It could be the objective of building a bridge, flying to the moon, developing new products, or even implementing a project. In a mobile world, it is important to have sets of marketing objectives available when we decide to apply CRM. Table 3 summarizes the relationship between marketing objectives and mobile CRM implementation to acquire and strengthen relationships with consumers.

Table 3 shows the six key marketing objectives of businesses and examples of CRM applications that enable development of one customer at one time.

Table 3. Relationship between marketing objectives and mobile CRM application

Marketing Objectives	Description	Mobile CRM Application for Amy
Awareness and brand involvement.	Create brand and product reach and awareness by encouraging consumers to interact with the brand to strengthen the benefit and emotional feel of consumers to a product.	Mobile and interactive TV convergence, voting, and program contest via SMS.
Sales leads generation and drive physical store traffic.	Generate subscription for mobile content, opt-in and drive customers to a retail store.	Dynamic offering by sending SMS promotion to Amy's mobile when she is in a shopping mall.
M-commerce.	Similar to e-commerce, the main objective is creating sales.	M-commerce application such as M-Ticketing.
Self-service/productivity.	Improve work productivity, reduce service request, and streamline process.	Property industry-specific mobile sales force automation.
Content and communities	Deliver premium mobile content to subscribers. Personalization and content freshness are keys for success.	Astrology content delivered via SMS. It includes horoscope, games that relate to Amy's horoscope, and movie star news that shares the same horoscope with Amy.
Customer retention	Retain customers and build long-term relationship for cross-selling opportunities.	Bank interest rate, forex update, account balance, SMS alerts, etc.

Figure 6. Mobile CRM, and integrated approach



THE INTEGRATION APPROACH

We have shown readers mobile CRM examples. To be successful in reaching, acquiring, converting, and retaining mobile consumers, we recommend that businesses align strategy, marketing objectives, and the customer engagement process with the appropriate technical platform and application. Figure 6 summarizes an integrated CRM approach in winning loyal consumers in a wireless world.

The spiral model in Figure 6 shows an abstract of combining strategy, marketing objectives, process, implementation, and measurement for a mobile CRM project. It is strongly believed that there is no priority on whether the entire marketing strategy drives the creation of mobile CRM campaign or technical design since all the proposed elements are equally important. For example, you might have a great vision of using a mobile device to acquire customers; however in terms of technical aspect, it would be not realistic or vice versa.

Furthermore, the use of a multi-channel approach and integration of different technology are also crucial for reaching, acquiring, and retaining wireless consumers. A mobile CRM

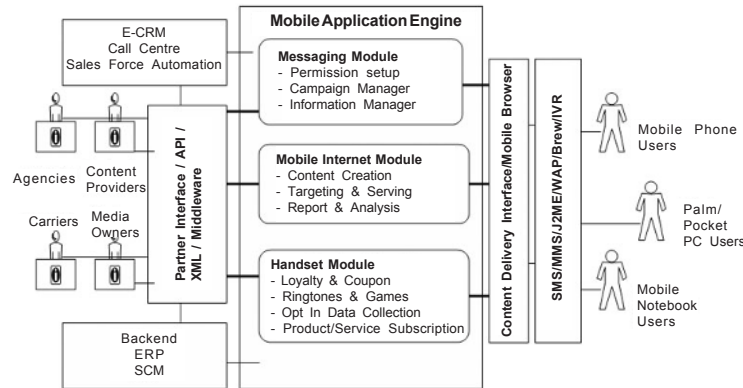
application needs to be integrated with an application server, middleware, partners' interface, and an other Web-based CRM and call centre system. Figure 7 shows the mobile CRM architecture, which is based on a leading mobile marketing solution vendor, Enpocket's (2005) mobile engine.

Based on the proposed mobile CRM architecture, end users are connected and receive mobile content using standards such as SMS, MMS, J2ME, and so forth. The core components that sit in the application engine are:

- **Messaging Module:** This module handles functions of setting up permission for campaign delivery, as well as a campaign manager that enables SMS blasting services and an information manager.
- **Mobile Internet Module:** This module deals with mobile content production, for example, astrology, news, sports updates, and other mobile content. It consists of content creation, the targeting function to match relevant content for end users, as well as reporting analysis to measure the overall content delivery performance such as number of subscriptions, content viewing, impression, response rate, and so on.
- **Handset Module:** This module deals with B2C-based services. The m-coupon redeems and loyalty-based sweep takers where we received in promotional leaflets lies in this module. It also functions as a download engine for ring tones and mobile network games. The interaction between user registration and games participation activity data collection also sits in this module.

Finally, the mobile application engine also needs to be integrated with backend systems and other CRM systems to support all CRM activities, from marketing campaign management to sales closing and after-sales services.

Figure 7. Mobile CRM architecture



CONCLUSION AND FUTURE PREDICTION

In conclusion, businesses and organizations are still facing obstacles in acquiring customers and delivering consistency data services to wireless consumers considering many possibilities and promises on mobile technology. Successful mobile projects need to identify the challenges of a wireless world such as screen size, bandwidth, usability, security, and content relevance. Therefore, segmenting the consumers, choose a positioning strategy, and effectively using the viral marketing concept are keys for success. In addition, the understanding of multi-channel conversion and matching CRM campaign design with marketing objectives is equally important. In a final word, we can build a critical mass, but establish one customer at a time in the mobile venture.

What are the implications of a wireless world in five years' time? I can say they are full of surprises and hopes. The following are some of the predictions:

- The boundary between carriers, agencies, content providers, and mobile marketers will become blurring in the future. Major

mergers and acquisitions between mobile service providers as well as smart strategic alliances cause this.

- New mobile solution players will emerge in the market. They will focus on delivering niche applications such as real estate management and map-based systems to vertical industry.
- A consumer-driven process means that consumers will require on-demand mobile data services and content that matches their needs and daily lives.
- The availability of new technology and delivery of promises will be important, for example, the complete 3G infrastructure will create a high bandwidth user experience where mobile devices will function as entertainment gadgets in receiving high-quality broadcast movies and live sports events.
- An emerging market is that of mobile teens and children.

Telstra, Australia's largest telecommunication company, has just introduced iMode services to Australian mobile users. iMode guarantees an optimistic future of mobile businesses, and will soon shape and improve pro-

ductivity for our work lives and increase fun in business and entertainment.

REFERENCES

Accenture. (2001a). *Moving customers: Building customer relationships and generating demand in the mobile data world*. Accenture.

Accenture. (2001b). *In lieu of interactive skin: Getting a grip on mobile commerce in the entertainment industry*. Accenture.

Bughin, J. R. (2004). Using mobile phones to boost TV ratings. *The McKinsey Quarterly*.

Enpocket. (n.d.). Retrieved from www.enpocket.com

Lindgren, M., Jedbratt, J., & Svensson, E. (2002). *Beyond mobile: People, communications and marketing in a mobilized world*. New York: Palgrave.

Sterne, J. (2002). *Web metrics: Proven methods for measuring Web site success*. New York: John Wiley & Sons.

Chapter XLVI

Factors Influencing Segmentation and Demographics of Mobile–Customers

Anne-Marie Ranft

University of Technology, Australia

ABSTRACT

This chapter addresses important factors for consideration when readying a mobile commerce business for global business, addressing both regional differentiation in demographics that influence classifications of customer segments, and differentiation in demographics within a region. Globally, not all customer segments have regular access to mobile commerce facilities, and even for those that do, other demographic factors can impede their potential as mobile-customers. When starting from an Anglo-centric perspective, it is vital to have awareness of global differences in culture, language, payment options, time zones, legal restrictions, infrastructures, product needs, and market growth that could either improve or inhibit mobile-customer uptake, and in the worst case, result in unexpected litigation.

INTRODUCTION

Mobile-customers should be considered as one of the most significant elements for a mobile commerce enterprise. Mobile-customers of the enterprise are those customers that use mobile devices—the most common ones being mobile phones, personal digital assistants (PDAs), and notebook PCs. Mobile commerce products can include: physical devices, applications, and ac-

cessories; access to the mobile infrastructure; and unrelated products and services marketed, bought, and sold using a mobile device as the communication tool.

Internet-based e-commerce interactions are generally categorised by the broad segments of Consumer (C), Business (B), and Government (G), and then decomposed into the relevant market segments. However, when undertaking global commerce, regional factors providing

Table 1. Potential access to mobile commerce and e-commerce—summary

	OECD 1990	1998/2000	Non-OECD 1990	1998/2000
Fixed and mobile telecommunications access paths per 100 inhabitants	41.1	72.1	2.7	7.8
Internet hosts per 1,000 inhabitants	23	82	0.21	0.85
Data source: OECD, 2001				
OECD Countries —There are 30 member countries, mainly in the European and North American regions, as well as the United Kingdom, Australia, and New Zealand.				

differentiation in demographics can alter classifications of customer segments, and differences in demographics can occur within a region. A market segment that exists in Australia, the United States, or the United Kingdom may not exist in some regions. It should also be noted that market segments based on Internet e-customer demographics may not necessarily be directly applicable for mobile-customers.

Before targeting a product or service to a particular market segment and location, these issues should be considered to maximise mobile-customer uptake and prevent unexpected litigation.

FACTORS INFLUENCING GLOBAL DIFFERENCES

First, the question of regional mobile-customer segment sizes will be discussed with reference to the *digital divide*, then other differentiating factors will be listed, followed by a list of possible strategies to consider when designing global mobile commerce products and marketing.

Digital Divide—Historical Factors

The first issue to be addressed is one of whether potential mobile-customers for a segment even exist in the targeted regions.

“Visions of a global knowledge-based economy and universal electronic commerce, characterised by the ‘death of distance’ must be tempered by the reality that half the world’s population has never made a telephone call, much less accessed the Internet” is the caveat noted by the Organisation for Economic Cooperation and Development (OECD, 1999). The OECD uses the term “digital divide” to describe the inter- and intra-country inequalities in access to information and communication technologies by both individuals and businesses due to socio-economic and geographic differences (OECD, 2001). They provided statistics that highlight the differences between OECD and non-OECD countries (see Table 1).

They further noted that the higher growth rate in telecommunication access for non-OECD countries is especially due to rises in China, but there was insignificant African growth during that period.

Within a geographic region, different demographic factors also contribute to a reduction in potential mobile-customers.

Uptake of mobile commerce in some regions is still biased towards the business and professional consumer sectors, especially mobile phone ownership in the Asian region.

It should be noted that many *developing* nations suffer from lack of suitable telecommunication infrastructure; access to a reliable

electrical source for re-charging of mobile devices and permanent housing can be limited for lower socio-economic groups. For instance, in my experience I have found that many Indian businesses have access to high-speed Internet lines and mobile connectivity, but require their own generators to back up the state power supply.

Overall, it can be concluded that there are two major groups of potential mobile-customer segments not currently available due to this *digital divide* factor—consumer and business segments whose geographical demographics are characterised by lack of telecommunication and other infrastructure, and consumer and business segments whose socio-economic demographics make mobile commerce unviable.

In many regions, especially Asia and Africa, consumer and small business sectors in lower socio-economic groups have the double barriers of no infrastructure and un-affordability, with the result that much of their population cannot today be counted as potential mobile-customers for C2C, B2C, and G2C segments.

Digital Divide—Transition Factors

The last few years have seen an enormous increase in the number of mobile phone connections in all global regions. This is shown in Table 2.

A common trend noted globally is the increase in the proportion of mobile subscribers to fixed telephone line customers. Some customer segments, especially youth segments in rental accommodation, may no longer see the necessity for a fixed line. Logistically, the resources required for installation of new mobile infrastructures in rural or undeveloped regions may be less than that required for new fixed-line infrastructures.

In Australia, the Australian Communications Authority's "Telecommunications Performance Report 2003-04" tabled that the number of mobile phone services had exceeded the number of fixed telephone services operating by June 2004. The number of mobile phone services grew by 15.4% over the period, with a growth in prepaid services, which by then made

Table 2. Mobile phone connections—summary

Region	1998 (1000s)	2003 (1000s)	CAGR (%) 1998-03 ^a	Per 100 Inhabitants 2003	As % of Total Telephone Subscribers 2003
Africa	4,156.9	50,803.2	65.0	6.16	67.3
Americas	95,066.8	288,219.9	24.8	33.80	49.8
Asia ^b	108,320.6	543,153.4	38.1	15.03	52.4
Europe ^c	104,382.0	441,234.9	33.4	55.40	57.5
Oceania	5,748.5	17,256.3	24.6	54.45	57.2
World	317,674.8	1,340,667.7	33.4	21.91	53.9
Notes: ^a The compound annual growth rate (CAGR) is computed by the formula: $[(P_v / P_0)^{(1/n)}] - 1$ where P_v = Present value P_0 = Beginning value n = Number of periods The result is multiplied by 100 to obtain a percentage. ^b By the end of 2003, Hong Kong and Taiwan had exceeded a rate of 100% phones per inhabitant. ^c By the end of 2003, Italy and Luxembourg had exceeded a rate of 100% phones per inhabitant. Data Source: International Telecommunication Union, 2004					

up 43% of mobile services (Australian Communications Authority, 2004).

Customer segments in regions with limited fixed-line infrastructures may now, for the first time, have access to modern telecommunications. Of particular interest is the increase in the size of the potential mobile-customer segments in regions that until 2002 were limited in the infrastructure required to support mobile commerce, thus enabling the creation of an emerging market segment.

The UN's International Telecommunications Union industry report, "Trends in Telecommunications Reform 2004-2005," has been reported by the press to state that globally, 2004 revenue from mobile services is expected to be higher than revenue from fixed telephone line services. China, India, and Russia were stated to have the highest rate of increase (Australian IT, 2004).

In India, a press release from the Telecom Regulatory Authority of India stated that during 2004, approximately 19.5 million mobile subscribers were added, giving a total of 48 million mobile subscribers (an increase of 68%). The number of mobile subscribers now exceeds that of fixed-line subscribers, who only experienced a small increase in numbers over the same period (Telecom Regulatory Authority of India, 2005).

Influence on Demographic Factors for This New Segment

The demographics of this emerging segment, especially of those located in less developed regions, may differ from early adopters of mobile commerce and Internet users in these regions by factors including:

- more likely to use a pre-paid account and less likely to own a credit card or have access to other e-commerce payment methods;

- mobile devices more likely to be limited to mobile phones, rather than business-oriented devices such as PDAs;
- wider geographic location—that is, rural areas without fixed-line telephony and Internet may now have access to mobile telephony infrastructure;
- wider age spread—that is, may be used for communication between many generations of a family structure; parents may purchase a mobile phone for their children to enable a sense of security, and conversely, adult children may purchase a mobile phone for their elderly parents to satisfy the same objective;
- may have attained lower levels of education and literacy;
- less likely to speak or read English, or even to be fluent in their own national language;
- may be less familiar with current communication technologies; and
- small businesses, especially in the rural sector, may now have access to mobile telephony, thus facilitating the potential for the deployment of new business and agricultural techniques.

Location Differentiation

Some differences affect all potential mobile-customers in a specific location, be it geographical region or individual country/province.

Geography

- Time differences in mobile-customers' time zones, established business hours, and public and religious holidays could affect peak and off-peak system processing loads, with implications for the scheduling of system downtimes for maintenance or upgrades, and the staffing of call centres and other customer services.

Factors Influencing Segmentation and Demographics of Mobile-Customers

- Seasonal and climate differences affect the marketability and usability of some products.
- Metropolitan vs. rural locality can impact the availability and quality of communications and product delivery infrastructure, unless the product can be delivered via the mobile device. Many Asian and African rural areas lack communication and other infrastructures, and even remote locations well serviced by satellite communications, such as the Australian outback or Antarctic bases, can have poor or expensive product delivery services.

Products and Services

- Suitability for use in global locations must be considered. Is there a need for the product or service? What use is a service to send payment details to a parking meter if few customers in the region own a car? Will the product actually work? This is especially an issue for electrical goods such as chargers for mobile devices or other items purchased via mobile commerce which may not be compatible with localequipment.
- Accuracy and knowledge of locality is important for some products, especially location-based services that interact with and require a global positioning system (GPS) infrastructure in the region.
- Social acceptance of products needs to be understood. Is the product attractive to the locations' typical mobile-customer needs, social values, and religious beliefs, or even legal?
- Equipment and availability for mobile commerce may differ for some customer segments. A business traveller expecting global availability of Wi-Fi "hotspots" for PDA or PC connection may be disap-

pointed when travelling in less technically developed regions, and there are some regions that are not yet reachable by commercial GPS satellites.

Handset types required depend on whether the local networks offer Global System for Mobile communication (GSM), Code-Division Multiple Access (CDMA) of which there are many variations, Personal Digital Cellular (PDC), or Third-Generation/Universal Mobile Telecommunications System (3G/UMTS) services. The Japanese network types are fairly unique to Japan; few commercially available handsets can be used both in Japan and other countries. Despite the availability of Japans' NTT DoCoMo iMode service in many countries, including Australia, the applications available and handsets required do differ between the individual countries of implementation.

- The number of potential customers who are visiting a region affect the viability of services that are aimed at the visitor, for example local directories, tourism guides, or special communication roaming deals such as SingTel's "Local Direct Dial" in Singapore (SingTel, 2005).

Product Content and Interface Presentation

- Language and keyboard/screen character sets differ. This is especially important to remember if mobile-customers are sought in China or Japan. Emerging mobile-customer segments may require mobile devices and applications to be designed using the local language for the interfacing component.
- Marketing promotions should be sensitive to customers' varying social backgrounds and local legislation regarding content.

Financial

- Credit card ownership is not ubiquitous in some regions. While customers can use other forms of payments such as invoicing, COD, or local debit cards for national purchases, credit cards are the most widely acceptable payment method for international mobile commerce. In many Western European countries, Spain and Germany in particular, most consumers use debit rather than credit cards, limiting their global mobile-customer potential (Barclays, 2001; Forrester Research Technology, 2004). Some Asian countries such as South Korea, Japan, and Hong Kong have high credit card ownership (Lafferty Cards International, 2004a), while most others do not.

On an optimistic note for global mobile commerce, data shows that credit card ownership is growing/is projected to grow strongly in the Asia Pacific (Visa, 2004a), especially Indian credit card use (Gupta & Dasgupta, 2004), as well as Central and Eastern Europe, and Middle East regions (Visa, 2004a). To overcome this limitation, billing options that integrate with the customers' mobile account should be considered, whether it is a pre-paid or post-paid account.

- Cash payments are preferred by consumers in some regions. Visa notes that over 90% of transactions in the Asia Pacific region are made in cash (Visa, 2004b). Some European countries such as Greece are still cash oriented (Lafferty Cards International, 2004b). Again, these customers could be catered for by billing options that integrate with the customers' mobile account, which may well be a pre-paid account.
- Currencies for transactions—can customers pay in their own currency, only in the

major currencies, or only in the currency of the mobile commerce business?

- Taxes—VAT, GST, state, and other sales taxes may or may not be payable on transactions depending on where the mobile commerce site is located and the location of the customer.

Legal

- Forbidden products both create and limit mobile commerce opportunities in some regions. There may be a large potential market for prohibited goods, especially in countries such as Saudi Arabia where alcohol and a range of other goods are forbidden (Department of Foreign Affairs and Trade, 2004) for a mobile commerce enterprise willing to engage in a high-risk venture. Otherwise, such products should not be included when targeting consumers in those regions to avoid causing offence, litigation, or censorship.

Mobile services and content that does not meet local legislative requirements could cause the loss of a mobile operator's license.

- Privacy regulations differ greatly across the world in regard to data collection and management, and unsolicited marketing. In Australia the Privacy Act applies to businesses with an annual turnover of more than \$3 million and all businesses of certain types (Office of the Federal Privacy Commissioner, 2005). There are no significant data protection laws in the U.S. at this point. Member countries in the European Union have some of the strictest data protection laws in the world which attempt to control their citizens' data stored in non-member countries too (European Commission, 2005).

Customer Differentiation

Within a location, individual customer demographic and lifestyle differences may alter the identification and classification of customer segments from standards in the mobile commerce's home location.

Demographic

- Age group usage may differ especially in locations where older groups have limited literacy. Younger groups may embrace internationalism and be confident using a wide range of services, including those marketed in the English language, while older groups may be more conservative and prefer using brands and services that reflect their own culture. Younger groups may be more confident using their mobile telephone for more than just telephony and are enthusiastic users of Short Message Service (SMS).
- Education is especially important in developing locations, where generally only the better educated have an opportunity to earn sufficient income to acquire the necessary infrastructure.
- Gender may affect customer segments in locations where females in lower socio-economic groups are less educated.
- Family lifecycle stage groups may differ in relative segment sizes. For instance, the relative size of the European "adult with no dependents" demographic is larger than that in many Asian countries.
- Metropolitan vs. rural locality differentiation is covered above. In some regions, education and financial infrastructures may also be limited in rural and remote areas.
- Language used may be different to the national language. Many regions comprise many ethnic language groups, espe-

cially India. English is more likely to be understood by the higher socio-economic groups.

Lifestyle

- Time consciousness—mobile devices are more likely to be used in the course of performing business functions when timing of communications is critical, or of a personal nature when the customer has limited time for family and social activities. Different cultures experience a difference in expectations of what is considered "on time" or not.
- Moral attitudes vary greatly, especially for sexuality. Various "adult" services of a sexual nature are marketed heavily to mobile customers in some regions, but could cause the loss of an operator's license if marketed or offered in a region with strict legislation controlling mobile content.
- Personal values differ between cultures, which should be taken into account when marketing and designing features. Is the target society one that values concepts of individuality or social and family group membership? Is there prestige associated with acquisition of new mobile and other technologies?
- Attitude to adoption of new technology may differ between different segments within a region. Japanese youth are well known for their enthusiastic embrace of mobile telephones, individualizing accessories, and mobile services offered in particular by their iMode system.

Firmographics

- Size does matter. Globally, smaller businesses are less likely to use the latest

technologies (OECD, 2001). Small businesses in developing locations are even less likely due to infrastructure issues listed above.

- Industry sector is shown to affect Internet use (OECD, 2001). Predominately subsistence-level agricultural communities may not require mobile commerce.

STRATEGIES FOR THE DESIGN OF PRODUCTS AND MARKETING

Strategies for the products, services, and marketing delivered by the mobile commerce business require tailoring for the targeted mobile-customer segments.

First, the customer segments should be identified according to the global differentiations outlined. Next, a decision should be made whether to create individual products, services, and marketing for different segments, or create a common suite to be used for all.

Factors indicating individual suites include:

- Significant differences in deliverable products and services, and customer differentiation, especially in legal restrictions, currencies, language, and social values.
- Economic justification for developing multiple products, services, and marketing campaigns.

Factors indicating a common suite include:

- Uniformity in products, services, and customer demographics.
- Uneconomic to develop multiple products, services, and marketing campaigns.

Then, the targeted mobile-customer segments should be guided to the appropriate site by strategies such as:

- Marketing and linking the mobile commerce product or service from an established mobile commerce portal, perhaps run by the telecommunication operator in the region or from a relevant Internet site in the region.
- Marketing via traditional channels such as print advertising in the region.

And finally, the mobile-customers should be provided with good “quality of service” regardless of their time zones and other differences. Consistent service availability and customer support should be provided to the most profitable customer segments at least, and ideally, to all.

CONCLUSION

While the recent arrival of mobile telecommunications infrastructures in most regions of the world has created a vast number of potential mobile-customers, mobile commerce businesses should be aware of the many geographical, legal, and demographic differences summarised in the following diagram before attempting to trade internationally, or deliver products and services developed outside their region to the local market.

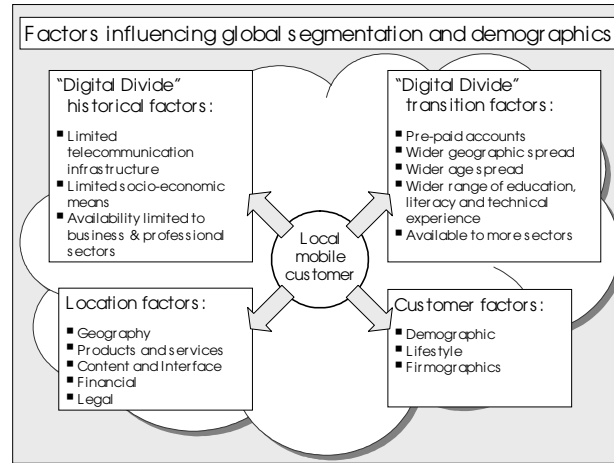
Shrinkage of the digital divide for business and medium-high socio-economic groups across international boundaries, especially in the Asian region and within developed countries, is enabling the potential for even more growth in the size and variety of mobile-customer segments.

The recent emergence of new potential mobile-customers outside the established socio-economic and urban-located demographic groups requires more careful tailoring of products, services, and billing options than for the more established segments.

Benefiting from this expected growth can only be achieved by ensuring the mobile

Factors Influencing Segmentation and Demographics of Mobile-Customers

Figure 1. Summary of influencing factors



commerce's products and services, interface design, and marketing; customer service is tailored to satisfy the targeted market segment, being either the established or emerging mobile-customer segments.

REFERENCES

- Australian Communications Authority. (2004, December). *Media release 95: Growth in mobiles and wireless broadband highlight year in telecommunications*. Retrieved January 22, 2005, from <http://internet.aca.gov.au>
- Australian IT. (2004, December 14). *Mobile revenue to outstrip landlines*. Retrieved January 10, 2005, from <http://www.australianit.news.com.au>
- Barclays. (2001). *International growth*. Retrieved September 10, 2004, from <http://www.investor.barclays.co.uk>
- Department of Foreign Affairs and Trade. (2004). *Department of Foreign Affairs and Trade, Saudi Arabia country brief*. Retrieved September 10, 2004, from <http://www.smartraveller.gov.au>
- European Commission. (2005). *Information society—Telecommunications, privacy protection*. Retrieved January 23, 2005, from <http://europa.eu.int>
- Forrester Research Technology. (2004, August). *Forrester's consumer technographics*. Retrieved September 10, 2004, from <http://www.forrester.com>
- Gupta, N. S., & Dasgupta, S. (2004). Dragon fire's no match for India's credit card club. *The Economic Times* (April 8). Retrieved September 10, 2004, from <http://economictimes.indiatimes.com>
- International Telecommunication Union. (2004). *Mobile cellular, subscribers per 100 people 2003*. Retrieved January 21, 2005, from <http://www.itu.int>
- Lafferty Cards International. (2004a, August). *Korean card use declines*. Retrieved September 10, 2004, from <http://www.lafferty.com>

Factors Influencing Segmentation and Demographics of Mobile-Customers

Lafferty Cards International. (2004b, August). *Olympian leap forward for Greek cards*. Retrieved September 10, 2004, from <http://www.lafferty.com>

OECD (Organisation For Economic Cooperation and Development). (1999). *The economic and social impact of electronic commerce: Preliminary findings and research agenda*. Retrieved September 11, 2004, from <http://www.oecd.org>

OECD. (2001). *Understanding the digital divide*. Retrieved September 11, 2004, from <http://www.oecd.org>

Office of the Federal Privacy Commissioner. (2005). *Private sector—business*. Retrieved

January 22, 2005, from <http://www.privacy.gov.au>

SingTel. (2005). *Visiting Singapore*. Retrieved January 23, 2005, from <http://home.singtel.com>

Telecom Regulatory Authority of India. (2005, January 9). *Press Release no. 6/2005*. Retrieved January 22, 2005, from <http://www.trai.gov.in>

Visa. (2004a). *Visa Asia Pacific*. Retrieved September 10, 2004, from <http://corporate.visa.com>

Visa. (2004b). *CEMEA*. Retrieved September 10, 2004, from <http://corporate.visa.com>

Section X

Social

Chapter XLVII

Mobile Camera Phones—Dealing with Privacy, Harassment, and Spying/Surveillance Concerns

Christopher Abood
Australian Computer Society, Australia

ABSTRACT

This chapter discusses the growing inappropriate use of mobile camera phones within our society. There are two areas of concern that are dealt within this chapter. The first concern deals with individual privacy and the use of mobile camera phones as a tool of harassment. The second concern deals with organizations seeking to prevent industrial espionage and employee protection. This chapter outlines how these devices are being used to invade individuals' privacy, to harass individuals, and to infiltrate organizations. The author outlines strategies and recommendations that both government and manufacturers of mobile camera phones can implement to better protect individual privacy, and policies that organizations can implement to help protect them from industrial espionage.

INTRODUCTION

During 2004, Samsung ran a television advertisement depicting a young man sitting in a café. He was taking photos of a young girl walking across the promenade with his mobile phone with in-built camera. The young girl was not aware that she was having her photo taken as she walked by. However, she turned to look over at the café and realized she was being

photographed. She walked over to the man sitting in the café, took the mobile camera phone from him, and began to take pictures of herself rolling over a car.

During the 2004 Olympics, LG ran a television advertisement where a girl on a beach phones her friend in a shop to show her a live video feed of a muscular man applying suntan lotion next to his surfboard. It is obvious that the man is unaware that he is being videoed. The

girl in the shop receiving the images passes the phone to a male companion to view. The catch phrase for the advertisement was: "LG, Official Sponsors of Eye Candy."

In both cases, these television advertisements depict voyeurism as a legitimate activity. In the Samsung advertisement, it goes further to suggest that people like having their photos taken with or without their knowledge or consent.

It is therefore ironic that both Samsung and LG have banned the use of mobile camera phones within their operations due to concerns these devices will be used for industrial espionage (BBC News, 2003). Also ironic is that Sydney, Australia, resident Peter Mackenzie was fined \$500 for using his mobile phone to photograph women sunbathing topless at Coogee Beach (The Australian, 2004). Appearing in Waverley Local Court, Mackenzie pleaded guilty to behaving offensively in a public place. He told reporters later that he regretted his actions and realized they had been inappropriate. However, his behaviour was actively encouraged by mobile phone vendor advertising campaigns. Mackenzie's actions on Coogee Beach were entirely consistent with advertising campaigns for mobile camera phone technology, but the fact that he was arrested, charged, and subsequently fined makes it clear that these campaigns are out of step with reality and public standards.

Also during 2004, Virgin ran a television advertisement depicting a game called Ming Mong. The game essentially involves sending a picture to someone's mobile phone with a caption. One such example in the television advertisement was a picture of a toilet with the caption, "your breath." This advertising campaign is out of step with community concerns over the increasing use of mobile phones to bully and intimidate others, especially within the school environment.

The past few years have seen rapid convergence within various technologies, none more so than the mobile phone. The mobile phone now has PDA functionality and the ability to send and receive e-mails, view Web pages, listen to the radio and MP3 songs, and play games. Mobile phones are now coming onto the market with one or more gigabytes of storage and of course digital camera facilities with 3-plus mega-pixel resolution. We are now starting to see the adoption of videoing facilities (enabling real-time chat), and it will not be long before these devices start incorporating global positioning system (GPS) mapping technologies (which raises all sorts of surveillance/stalking issues). In short, the mobile phone is morphing into the everyday must-have mobile information and communications centre.

This convergence in technology, while providing many benefits, also raises issues dealing with privacy and surveillance/spying. Although we have had digital photography for a number of years, people generally tend not to carry their digital cameras with them all the time, whereas people tend to carry their mobile phones with them constantly. With a digital camera, you need to go home, connect it to your PC, and transfer the images from the camera to a storage medium. You then possibly e-mailed the images or uploaded them to a Web site. However, with a mobile camera phone, you can immediately send the image to an e-mail address, another phone, or to a computer server (for display on a Web site).

This can be great if you have one of those photo magic moments that you wish your friends to share. But it is not so great if the photo being forwarded is one that has been taken without the subject's knowledge. It is not readily obvious if someone is using their mobile camera phone to take photographs, as they may appear to be just chatting on the phone. Most mobile camera phones have the lens on the back, so

when you are talking, the lens has a clear view to your side. Some mobile camera phones have a swivel lens, which makes it easier to conceal the fact that you are taking a photograph. You can now purchase a mobile camera phone relatively cheaply, walk into a competitor's premises, take photos of a sensitive industrial nature (while pretending to be talking to someone), immediately send on the image, dump the phone into a bin, and walk out clean.

Individual privacy and industrial espionage are becoming two major concerns dealing with mobile camera phones. Governments and organizations are grappling with how to deal with the growing misuse of these devices, and there does not seem to be a clear answer. How we protect an individual's and an organization's privacy while still allowing people to enjoy the benefits that mobile camera phones provide will be a difficult juggling act.

In this chapter, I discuss how to juggle protecting individuals' and organizations' rights, while at the same time enabling users to gain the benefits of mobile camera phones. Much of the discussion will be derived from the Australian Computer Society's policy on mobile camera phones (of which I led the development). This policy will be discussed at the end—but first, protecting individual privacy.

INDIVIDUAL PRIVACY

A number of Web sites have appeared that cater for images taken by mobile camera phones. There is www.mobog.com, which is a Web blog (an online journal) for images taken by mobile camera phones. People can create their own section to upload, store, and display photos taken with their mobile camera phone. This is particularly useful if you are travelling and wish friends back home to view where you have been. www.phonepiks.com is a central

place for people to upload and view pictures taken with their mobile camera phone. Unlike [mobog](http://www.mobog.com), photos are uploaded to predefined categories. Both these sites contain images that range from the relatively mundane through to the pornographic. Another site, www.mobileasses.com, specializes in images taken of people's backsides by mobile camera phones. The photos are displayed with information about the type of mobile phone used, where it was taken, and when. The site even has a "backside of the day" competition, where viewers vote on the picture of a backside they like best. The winner wins a t-shirt. In most instances of images involving people displayed on these sites, it is obvious that the picture has been taken without the subject's knowledge or consent. They are probably unaware that their picture resides on these sites, let alone what to do to have the image removed. Mobileasses.com even has a section on tips on how to take pictures of people's backsides without their knowledge.

It is inappropriate to take a photo of someone without his or her knowledge or consent. It is even more so to then upload that image where all control over the image is lost. Even if the victim were successful in having his or her image removed, it would be of little consequence, as the image can be re-downloaded and forwarded on many times. Such is the nature of the Internet. Once the image has left the geographical boundaries of a nation, an individual has little chance of having the image removed, especially as they will now have to deal with a number of different legal systems in various countries. The difficulty with dealing with different legal systems is highlighted by the recent legislation passed by the Australian government. It would be illegal for a site like mobileasses.com to be hosted in Australia. A complaint can be made and a take down notice issued by the Australian Broadcasting Agency.

This means the Web site must be removed from the server and access denied to the Web site (in Australia). Many such sites in Australia have been issued with such take down notices. All of these sites simply relocated to servers located in other countries where the Australian law has no jurisdiction and have continued business uninterrupted. These sites still use the same Web address as before with the .au domain.

Like e-mail in its early days, it is difficult to predict what other ways people will use mobile camera phones in the future. Examples include “up-skirting” (prevalent on mobileasses.com) and digital shoplifting. Up-skirting involves using your mobile camera phone to take photos up a lady’s skirt. On mobileasses.com, there are plenty of examples of a photograph being taken under a table. Holding a mobile camera phone in your hand looks inconspicuous and not readily obvious that your photo is being taken. Digital shoplifting involves someone taking a photo of an article in a magazine in a newsagent without having to buy the magazine. Some newsagents have banned the use of mobile camera phones from within their premises, but again, unless you are going to search each person as they enter and make them leave their phone at the counter (as some make you leave your bag at the counter), it is going to be hard to eliminate this practice.

Probably the most concerning of inappropriate use of mobile camera phones is using images taken of people in awkward situations for bullying and intimidation. Indeed, the biggest problem of using mobile phones for bullying is via the use of Short Message System (SMS). However, I believe that Multimedia Messaging Service (MMS) will replace SMS as the preferred means of bullying and harassment (as encouraged in the Virgin Ming Mong television advertisement). MMS will allow you to send images, text, video, and audio. So without your knowledge, your conversation can be recorded,

your movements videoed, and images taken of you. These can then be used against you, especially if what has been captured is of an embarrassing nature. The video, image, and audio that is captured can also be uploaded immediately to a Web site for the entire world to view. This is what happened to a 17-year-old Indian boy who used his mobile camera phone to record his girlfriend giving him oral sex (Sydney Morning Herald, 2004). The video clip somehow made its way onto the Internet. This is something that they probably did not want to happen.

So how can we protect our privacy and control over our image without infringing on the benefits that these devices provide, such as documenting an accident? Simply banning mobile camera phones from places such as beaches, swimming pools, gyms, and other public places will not work. Not all mobile phones have in-built cameras, and people have a genuine need to have their mobile phones with them. To effectively ban these devices, you would need to search everyone, which would be impractical. As stated earlier, it is not always obvious that someone talking on his or her mobile phone is actually taking a photograph. You may also open yourself up for legal action if you confront someone who is actually just talking on the phone.

Governments are also grappling with the problem of misuse by passing various laws; however, it is difficult to enforce these laws. In New South Wales, Australia, it is against the law to drive and talk on your mobile phone at the same time, but I see people driving and chatting everyday. The United States is looking at passing a law to make it an offence to photograph people in situations where they would expect to have a reasonable amount of privacy. So you can take their photo on a public beach (which Peter Mackenzie found out you cannot do in Australia), but not in their backyard. However,

I believe this law would be open to interpretation; for instance, if I take a photo of you in your backyard at a party, is it still reasonable to expect a high level of privacy?

Again, laws will not help if you see a picture of yourself on a Web site and you have no idea who took it or how you could track down the person who uploaded it to the Web site. The local law would be even less powerful if that Web site resides in a different country with its own set of laws.

I believe the best way to reduce the amount of misuse of mobile camera phones is through education. For starters, mobile camera phone manufacturers should stop advertising their products in a way that encourages owners to use mobile camera phones in an inappropriate (and sometimes illegal) way. A mobile camera phone etiquette guide should be developed by mobile camera phone manufacturers (with appropriate input from the community) to be distributed with all new mobile camera phones. A campaign should be undertaken to educate users as to their responsibilities and the appropriate use of mobile camera phones. Both government and manufacturer should undertake this campaign.

The community must also be fully informed about their rights and what they should do if they suspect that they have had their photo taken without their consent. When taking a photo, you should make it clear that you are doing so and seek permission of the person being photographed, especially if they are unknown to you. If you do take a photo, you should take all reasonable steps to ensure that the photo is not uploaded to a Web site or e-mailed to others without the photographed person's permission.

For public venues such as gyms and swimming pools, visible signs should be displayed that indicate use of mobile camera phones, indeed any camera, is prohibited. A sign de-

signed similar to the non-smoking sign would be ideal. People who visit these public places should be made fully aware of their responsibilities and any penalties that may apply if they are caught photographing within the premises. Penalties could include banishment or suspension from the premise.

It is hoped that when governments are passing laws into inappropriate use of mobile camera phones, and indeed any organization passing rules, that they do not infringe upon a person's right to use these devices legitimately. Mobile camera phones can enhance your safety and prevent crime. In Yokohama, Japan, an 18-year-old female store clerk used her camera phone to take a photo of a 38-year-old man who was fondling her on a commuter train. She called police during the train ride and presented her phone shots as evidence. The man was arrested at the next stop (CBS News, 2003).

It should be noted that if you suspect that someone has taken your photo without your consent, you should contact the appropriate authority. Try and avoid confronting the person unless you are sure that it is safe to do so; your physical safety is far more important.

As stated earlier, it is difficult dealing with Web sites that are hosted in a different country. Such is the nature that the Internet knows no borders. Perhaps it is time to look at an international approach to dealing with complaints against Web sites that publish images of you without your permission.

ORGANIZATIONAL PRIVACY

Business has also started to show concern regarding the use of these devices on their premises. Industrial espionage and spying is of real concern, namely because of the ease of use of mobile camera phones and the fact that it is not readily obvious that a person is captur-

Mobile Camera Phones

ing images while using them. It is difficult to tell whether a mobile phone is equipped with a built-in camera. A number of organizations will not issue staff with a mobile phone that has a built-in camera. Apart from spying and espionage concerns, there is the likelihood that if photos are taken of employees inappropriately either by employers or other employees, the organizations may face legal action for invasion of privacy or workplace harassment.

Mobiles with a built-in camera are not their organizations' only concern. Mobiles equipped with General Packet Radio Service (GPRS) may allow the downloading of data without your knowledge. Once downloaded, it can immediately be forwarded on to any number of destinations.

Raghu Raman of Mahindra Special Services Group says:

Some corporates are becoming sensitive that a camera phone is a potential snooping or espionage device and should not be treated lightly in high-risk areas. But most companies do not realize this and are blissfully unaware of leakage of sensitive information through this route. (India Times Infotech, 2004)

Some businesses have tried to deal with the problem by requiring sticky tape to be placed over the lens or banning mobiles from their premises. Neither of these options works. Sticky tape can easily be removed, and banning would require a search to be effective. It would also be difficult to ban mobile camera phones when employees need to be contactable, whether for business or personal reasons. There is also the prospect of using jamming technologies, but these are expensive and again would prevent vital communications, such as emergency calls. However, jamming technologies would be useful in sensitive areas.

Organizations will need to start looking at a variety of solutions to protect both sensitive information and personal privacy. Organizations will need to look at two separate and simultaneous sets of solutions, those that deal with employees and those that deal with non-employees interacting with your organization.

For employees, an organization should first begin with reviewing and incorporating acceptable use of mobile camera phones within their Human Resources policies. Employees should be made aware (on a regular basis) of the issues involved with using a mobile camera phone. It is an education step, the same as the education process that we went through for safe work practices and sexual harassment. Employees need to be made aware that taking a photo of a workplace colleague without their consent can constitute harassment. Taking a photo of corporate documents or downloading data without permission can constitute theft. Education of employees to their responsibilities will be important in avoiding future problems where employees may take legal action because of harassment and privacy violations.

Where companies issue an employee a mobile phone, they should have in place policies that limit those who are issued with a built-in camera only if it is required as part of their work. Many companies now only issue to employees mobile phones without built-in cameras. Samsung, manufacturer of mobile phones, stated that it had been asked by Telstra and Optus (two Australian communications companies) to continue to offer conventional phones to corporate clients.

The general manager of mobile phones for Samsung Australia, Josh Delgado, said:

We have spoken to our carrier customers, Optus and Telstra, who sell to the corporate market, and they have mentioned that camera phones and this area of privacy are an

issue. They would not mention who those customers were, but we've been told that we should continue to manufacture both. (Lee, 2004)

Employees who bring their own mobile phones to work need to be aware of appropriate use within the organization. It will be difficult and impractical to ban private mobiles from the workplace, as people often need to be contactable.

Organizations that have non-employees interacting within their premises need to look at other solutions to protect sensitive information and personal privacy. As stated earlier, requiring sticky tape to be placed over the lens or banning them from their premises would be counterproductive, as sticky tape can easily be removed and banning would require a search to be effective. Also, these people would need their mobile phones to contact others, especially if they have to contact someone from their own organization for information or advice.

Organizations can look at requiring all entities interacting with the organization and be required to sign a non-disclosure agreement—that is, that all images and data of an organization remain the copyright of the organization. The agreement should also state that they would not use their mobile camera phone to take photos or download data without prior written permission. The agreement should also outline penalties that will be levied if the agreement is broken. This can be incorporated into existing agreements that your organization may already require. Again, making non-employees aware of the issues will greatly help in avoiding future problems.

Organizations need to also assess future changes in this technology and the impact this will have. One such change likely to come about is the incorporation of GPS mapping technologies within mobile phones. This will

provide enormous benefits to organizations and their employees. Employees will be better able to find clients' premises, and employers who need to deploy staff to a client's premises (for example, their computer has gone down) will be able to find the closest employee easily and readily.

However, organizations and employees need to be aware that privacy issues can arise. For instance, does an organization tracking an employee outside of business hours, or during breaks, constitute a privacy invasion? Also, tracking employees with the view to enhance productivity may expose the organization to potential legal problems. For example, an organization tracks a courier to ensure that they are not taking an excessive amount of time to make individual deliveries. If the courier knows that they are being monitored and can face disciplinary action for taking too long, they may start to take risks to shorten their delivery times to avoid disciplinary action. If these risk lead to an accident, the organization may find that they are liable if it can be shown that the monitoring led to the risk to shorten the delivery time.

The point is, when an organization implements the use or allows the use of new technologies such as camera phones or GPS-equipped phones, they need to think through the implications of these technologies and how people may use them. Clear guidelines need to be established, and employer and employee education needs to be undertaken to ensure that the organization's policies and guidelines are being met.

AUSTRALIAN COMPUTER SOCIETY POLICY ON MOBILE CAMERA PHONES

The Australian Computer Society has taken a leadership position on this topic with the release

of its mobile camera phone policy in July 2004 (Abood, 2004). The policy has been well received, with both government and the media showing much interest in it.

Many have asked me why the Australian Computer Society is getting involved in this subject. Information and communication technologies have become prevalent within our lives. From school to work to home, we are becoming more reliant on these technologies. The benefits these technologies are providing us are many. However, these technologies also can be used to the detriment of others. As the Australian Computer Society is the guardian of ethics within our industry in Australia, it is our responsibility to alert society to inappropriate use of these technologies, and help guide society to better understand and use these technologies appropriately.

At the time that the policy was released, there was (and still is) a lack of clarity with regards to the protocols, policies, and laws regarding the use and misuse of camera phones. The policy recommends the following:

- Raising the level of awareness and potential impact of camera phones with Australian companies and the Australian public, including an education campaign to advise of rights, guidelines, and etiquette on the appropriate use of mobile phone cameras.
- Assessing the full implications of this technology, including relevant overseas developments.
- Introducing appropriate guidelines by businesses to address use of phone cameras as a “tool of trade”.
- Developing responsible guidelines by manufacturers, retailers, and promoters of this technology to be distributed with all phone cameras sold/issued and to accompany advertising of these products.
- Examining the use of technological solutions to enhance privacy including the use

of sound or a flashing light when a picture or video is being taken and the use of jamming technology.

- Developing a code of conduct for the use of mobile camera phones in sensitive areas such as change rooms.

The Australian Computer Society is also calling for a more transparent debate on this topic among the Australian federal and state governments and the telecommunications industry. This debate should be inclusive to ensure policy making in this area is not influenced solely by those with a commercial vested interest. Greater clarity and consistency in the approach taken by government(s), and by Australian corporations and venues to policy making in this area need to be undertaken.

The Australian Computer Society is also calling for greater awareness and education regarding:

- The protocols and legal responsibilities of the holders/users of this technology.
- The potential for misuse of this increasingly pervasive technology.
- The civil rights and protections that are in place should a member of the public feel they have been subject to abuse or misuse of this technology.
- The appropriate security precautions Australian companies can take in relation to this new technology.
- The appropriate use of this technology in public advertising—in light of the personal privacy concerns.

CONCLUSION AND FUTURE DIRECTION

While mobile camera phones provide many benefits, these devices are increasingly being used for inappropriate use. Areas of inappro-

priate use include invasion of privacy and use as a tool for harassment and spying/surveillance within organizations. With appropriate policies and education of the public with regards to the appropriate use of mobile camera phones, we will be all able to enjoy the benefits of this technology without the unwanted side effects.

REFERENCES

- Abood, C. (2004, July). *The Australian Computer Society policy on mobile camera phones*. Retrieved from http://www.acs.org.au/acs_policies/docs/2004/mobilespolicyfinal.pdf
- The Australian*. (2004). Fine over mobile topless pictures. *The Australian*, (December 1). Retrieved from http://www.theaustralian.news.com.au/commonstory_page/0,5744,11555056%255E1702,00.html
- BBC News. (2003, July 7). *Samsung bans "spy" phones*. Retrieved from <http://news.bbc.co.uk/1/hi/world/asia-pacific/3052156.stm>
- CBS News. (2003, July 9). *Camera phone etiquette abuses*. Retrieved from <http://www.cbsnews.com/stories/2003/07/09/tech/main562434.shtml>
- India Times Infotech. (2004, December 23). *Camera phone: A security threat*. Retrieved from <http://infotech.indiatimes.com/articleshow/968730.cms>
- Lee, J. (2004). Is that a camera in your briefcase? *The Sydney Morning Herald* (June 12). Retrieved from <http://www.smh.com.au/articles/2004/06/11/1086749894185.html?from=storylhs&oneclick=true>
- The Sydney Morning Herald*. (2004). Indian schoolboy's phone sex prank reverberates around the world. *The Sydney Morning Herald* (December 22). Retrieved from <http://www.smh.com.au/news/World/Indian-school-boys-phone-sex-prank-reverberates-around-the-world/2004/12/21/1103391774573.html?oneclick=true>

Chapter XLVIII

Social Context for Mobile Computing Device Adoption and Diffusion: A Proposed Research Model and Key Research Issues

Andrew P. Ciganek

University of Wisconsin-Milwaukee, USA

K. Ramamurthy

University of Wisconsin-Milwaukee, USA

ABSTRACT

The purpose of this chapter is to explore and suggest how perceptions of the social context of an organization moderate the usage of an innovative technology. We propose a research model that is strongly grounded in theory and offer a number of associated propositions that can be used to investigate adoption and diffusion of mobile computing devices for business-to-business (B2B) interactions (including transactions and other informational exchanges). Mobile computing devices for B2B are treated as a technological innovation. An extension of existing adoption and diffusion models by considering the social contextual factors is necessary and appropriate in light of the fact that various aspects of the social context have been generally cited to be important in the introduction of new technologies. In particular, a micro-level analysis of this phenomenon for the introduction of new technologies is not common. Since the technological innovation that is considered here is very much in its nascent stages there may not as yet be a large body of users in a B2B context. Therefore, this provides a rich opportunity to conduct academic research. We expect this chapter to sow the seeds for extensive empirical research in the future.

INTRODUCTION

What causes individuals to adopt new information technologies (ITs)? How much influence do the perceptions of the social context of an organization have on the acceptance of new ITs? These questions are significant because systems that are not utilized will not result in expected efficiency and effectiveness gains (Agarwal & Prasad, 1999), and will end up as unproductive use of organizational resources. Academic research consequently has focused on the determinants of computer technology acceptance and utilization among users. Some of this research comes from the literature on adoption and diffusion of innovations (DOI), where an individual's perceptions about an innovation's attributes (e.g., compatibility, complexity, relative advantage, trialability, visibility) are posited to influence adoption behavior (Moore & Benbasat, 1991; Rogers, 2003). Another stream of research stems from the technology acceptance model (TAM), which has become widely accepted among IS researchers because of its parsimony and empirical support (Agarwal & Prasad, 1999; Davis, 1989; Davis, Bagozzi, & Warshaw, 1989; Hu, Chau, Sheng, & Tam, 1999; Jackson, Chow, & Leitch, 1997; Mathieson, 1991; Taylor & Todd, 1995; Venkatesh, 1999, 2000; Venkatesh & Davis, 1996, 2000; Venkatesh & Morris, 2000).

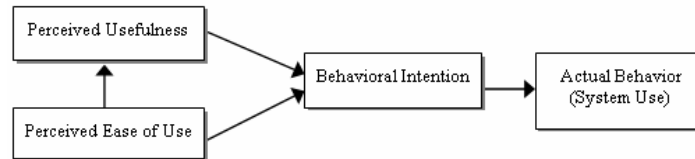
Individual differences indeed are believed to be very relevant to information system (IS) success (Zmud, 1979). Nelson (1990) also acknowledged the importance of individual differences in affecting the acceptance of new technologies. A variety of research has investigated differences in the perceptions of individuals when using TAM (Harrison & Rainer, 1992; Jackson et al., 1997; Venkatesh, 1999, 2000; Venkatesh & Morris, 2000); however, the perceptions and influences of the social context of an organization have not been widely examined in the literature. Hartwick and Barki (1994) suggest

that it is imperative to examine the acceptance of new technologies with different user populations in different organizational contexts.

Although mobile computing devices have existed for several years, strategic applications of this technology are still in their infancy. Mobile computing devices (in the context of business-to-business—B2B) is treated as a technology innovation in this chapter due to their newness and short history. An investigation into the usage of mobile computing devices within a B2B context, which we define as two or more entities engaged within a business relationship, is of value because of its increasing popularity (March, Hevner, & Ram, 2000). As an emergent phenomenon, relatively modest academic literature has examined the nature of adoption and use of this technology. Mobile computing devices, which have been described as both ubiquitous (March et al., 2000) and nomadic (Lyytinen & Yoo, 2002a, 2002b), offer a stark difference from traditional, static computing environments. A good characterization of these differences is provided in Satyanarayanan (1996). New technology innovations typically require changes in users' existing operating procedures, knowledge bases, or organizational relationships (Van de Ven, 1986). Such innovations may even require users to develop new ways of classifying, examining, and understanding problems. The domain of mobile computing devices has the potential to become the dominant paradigm for future computing applications (March et al., 2000), and topics of such contemporary interest are recommended to be pursued in IS research (Benbasat & Zmud, 1999; Lyytinen, 1999).

The primary objective of this chapter is to examine whether and how perceptions of the social context of an organization moderate the adoption, use, and infusion¹ of mobile computing devices for B2B transactions. We extend TAM to include individuals' perceptions of the social context of their organization, which in-

Figure 1. Technology acceptance model (Adapted from Davis, Bagozzi, & Warshaw, 1989)



corporate aspects of both culture and climate research as recommended in the literature (Denison, 1996; Moran & Volkwein, 1992). Aspects of the social context of an organization are suggested as having a significant role in the introduction of new technologies (Boudreau, Loch, Robey, & Straub, 1998; Denison & Mishra, 1995; Legler & Reischl, 2003; Orlikowski, 1993; Zammuto & O'Connor, 1992), particularly with the introduction of mobile computing devices (Jessup & Robey, 2002; Sarker & Wells, 2003). Only a handful of studies in the past have specifically looked at the micro-level connections of these relationships (Straub, 1994); unfortunately, even this has not been within a mobile computing context. We argue that an organization's social context will have a significant moderating effect on the perceptions of employees considering adoption and use of mobile computing applications for B2B purposes.

The chapter proceeds as follows: the next section presents the background research in the domains (adoption and diffusion of technology innovations within the context of TAM, DOI, and social context) underlying this research. This will be followed by the presentation and discussion of our proposed model and accompanying propositions. A brief discussion of the types of B2B application domains that are relevant to mobile-computing and would be of (future) interest to our investigation is then presented, accompanied by one methodological approach to how such research can be conducted. This chapter concludes with some po-

tential implications for research and practice, limitations of the book chapter, and potential future directions.

BACKGROUND RESEARCH

In this section, we first discuss the extant research connected with the technology acceptance model followed by research related to social context.

Technology Acceptance Model

The technology acceptance model proposed by Davis (1989) has its roots in the theory of reasoned action (TRA) of Fishbein and Ajzen (1975). As earlier alluded to, it is one of the most widely used models of IT acceptance. This model accounts for the psychological factors that influence user acceptance, adoption, and usage behavior of new IT (Davis, 1989; Davis et al., 1989; Hu et al., 1999; Mathieson, 1991; Taylor & Todd, 1995). The TAM model is displayed in Figure 1.

As is fairly well known in the IT literature, TAM specifies two beliefs—*perceived usefulness* (PU) and *perceived ease of use* (PEOU)—to be determinants of IT usage. It incorporates behavioral intention as a mediating variable in the model, which is important for both substantive and sensible reasons. In terms of substantive reasons, the formation of an intention to carry out a behavior is thought to be a necessary precursor to actual behavior

(Fishbein & Ajzen, 1975). In terms of sensible reasons, the inclusion of intention is found to increase the predictive power of models such as TAM and TRA, relative to models that do not include intention (Fishbein & Ajzen, 1975). Perceived usefulness is defined as “the degree to which a person believes that using a particular system would enhance her/his job performance”; perceived ease of use is defined as “the degree to which a person believes that using a particular system would be free of effort” (Davis, 1989, p. 320).

The TAM model and other subsequent IT models of acceptance have largely ignored the influence that continued usage has on the acceptance of an IT. For example, Karahanna, Straub, and Chervany (1999) found differences in the determinants of attitudes between potential adopters and actual users of an IT. In particular, they found that perceived usefulness continued to play an important role in the attitudes of IT users, while ease of use ceased to be important over time. Consequently, the relationship between actual/demonstrated usefulness and continued use is added by us to the original TAM model. Once the actual/realized usefulness of an IT is confirmed by a potential adopter, it is likely to continue to play a significant role in the overall infusion of the technology.

Based upon conceptual and empirical similarities across eight prominent models in the user acceptance literature, Venkatesh, Morris, Davis, and Davis (2003) developed a unified theory of individual acceptance of technology (the unified theory of acceptance and use of technology, or UTAUT). The UTAUT theorizes four constructs as having a significant role as direct determinants of acceptance and usage behavior: performance expectancy (subsuming perceived usefulness), effort expectancy (subsuming perceived ease of use), social influence, and facilitating conditions. In addition, it considers four moderators—age, gen-

der, voluntariness of use, and experience of the users to influence the relationship between the four direct antecedents and intentions to use (and in the case of facilitating conditions on actual use behavior). Although the UTAUT model explains a significant amount of variance in the intention to adopt an IT, the model lacks the parsimony and empirical replication of the TAM model. In this light, the modified TAM model that we propose may be considered a viable and prudent alternative to the UTAUT model. An empirical comparison between these two models is, of course, necessary.

Recent research employing the TAM model had identified individual differences as a major external variable (Agarwal & Prasad, 1999; Jackson et al., 1997; Venkatesh, 2000; Venkatesh & Morris, 2000). Individual differences are any forms of dissimilarity across people, including differences in perceptions and behavior (Agarwal & Prasad, 1999). For example, Agarwal and Prasad (1999) found that an individual's role (provider or user) with regard to a technology innovation, level of education, and previous experiences with similar technology were significantly related to their beliefs about the ease of use of a technology innovation. Agarwal and Prasad also found a significant relationship between an individual's participation in training and their beliefs about the usefulness of a technology innovation. Jackson et al. (1997) examined variables such as situational involvement, intrinsic involvement, and prior use of IT by users, and Venkatesh (2000) considered individual specific variables such as beliefs about computers and computer usage, and beliefs shaped by experiences with the technology in the traditional TAM. Both these studies found significant relationships among these individual differences and TAM constructs. Further, Venkatesh and Morris (2000) argue from their findings that “men are more driven by instrumental factors (i.e., per-

ceived usefulness) while women are more motivated by process (perceived ease of use) and social (subjective norm) factors” (p. 129). Thus, while the various above-noted research studies have investigated the differences in the perceptions of individuals using TAM as the underlying theoretical basis, as noted earlier, perceptions of the social context of an organization is not common in the literature. Most of these refinements to TAM and findings are accommodated in the earlier-noted overarching UTAUT model proposed by Venkatesh et al. (2003).

Social Context of an Organization and Innovativeness

As noted in the introduction, although the social context of an organization has been suggested as having a significant role in the introduction of new technologies (Boudreau et al., 1998; Denison & Mishra, 1995; Legler & Reischl, 2003; Orlikowski, 1993; Zammuto & O'Connor, 1992), particularly with the introduction of mobile computing devices (Jessup & Robey, 2002; Sarker & Wells, 2003), it has not been widely examined in the literature. In this chapter we extend the TAM to incorporate an individual's perceptions of the social context of their organization. The perceptions of the social context are of value to consider since they are likely to be fairly stable in the mind of the potential adopter and less subject to change than other perceived factors or the underlying technological innovation. As recommended in the literature, we examine the social context of an organization to incorporate aspects of both culture and climate (Denison, 1996; Moran & Volkwein, 1992). We take the stand that a study of organizational culture and organizational climate actually examine the same phenomenon—namely, the creation and influence of social contexts in organizations—but from

different perspectives (Denison, 1996). Following the recommendation of prior research, we examine the broader social context in order to improve our understanding of the organizational phenomenon (Astley & Van de Ven, 1983; Denison, 1996; Moran & Volkwein, 1992; Pfeffer, 1982).

Organizational climate can be described as the shared perceptions of organizational members who are exposed to the same organizational structure (Schneider, 1990). Zmud (1982) suggests that it is not the structure of the organization that triggers innovation; rather, innovation emerges from the organizational climate within which members recognize the desirability of innovation, and within which opportunities for innovation arise and efforts toward innovation are supported. As summarized in Schneider (1990) and in Moran and Volkwein (1992), a number of different conceptualizations of organizational climate have been suggested over the years. Pareek (1987) advanced the idea that climate and culture can only be discussed in terms of how it is perceived and felt by individual members/employees of the organization, which is a perspective that is supported in the literature (Legler & Reischl, 2003). Thus, we are interested in capturing the perceptions of individuals within organizations. Since the unit of analysis (during empirical evaluation) in this chapter is the individual employees within organizations, appropriate measures of examining social context can be derived from psychological climate literature.

Rather than focusing on how the psychological climate of an organization gets formed and can be influenced (certainly important), of interest in this chapter is how the prevailing climate of an organization moderates the relationship between individuals' perceptions of an innovation's usefulness and ease of use, and their intentions to adopt and use the innovation. Psychological climate is a multi-dimensional

Table 1. Dimensions of psychological climate (Adapted from Koys & DeCotiis, 1991)

Dimension Name	Definition
<i>Autonomy</i>	Employee's perception of their own sovereignty with respect to work procedures, goals and priorities.
<i>Cohesion</i>	Employee's perception of sharing and togetherness within their organization.
<i>Trust</i>	Employee's perception of freedom to communicate openly with members at higher organizational levels about sensitive or personal issues with the expectation that the integrity of such communications will not be violated.
<i>Pressure</i>	Employee's perception that time demands are incongruent with respect to task completion and performance standards.
<i>Support</i>	Employee's perception of the tolerance of their behavior by superiors, including the willingness to let employees learn from their mistakes without fear of reprisal.
<i>Recognition</i>	Employee's perception that their contributions to their organization are acknowledged.
<i>Fairness</i>	Employees' perception that their organization's practices are equitable and non-arbitrary.
<i>Innovation</i>	Employee's perception that change and originality are encouraged and valued within their organization, including risk-taking in domains where the individual may have little to no prior experience.

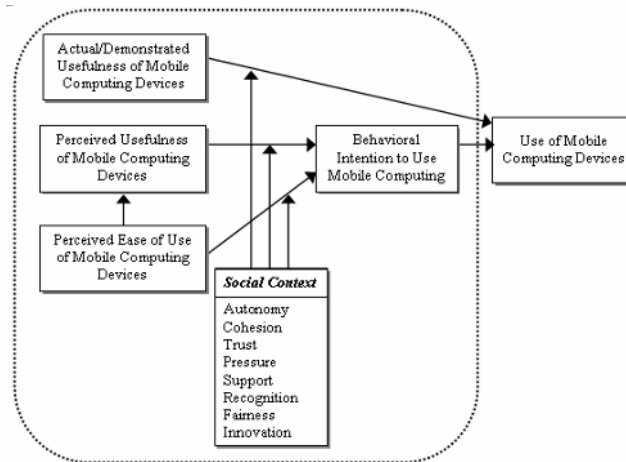
construct that can be conceptualized and operationalized at the individual level (Glick, 1985; Legler & Reischl, 2003). In an attempt to integrate several different measures of psychological climate, Koys and DeCotiis (1991) derived eight summary dimensions—*autonomy*, *cohesiveness*, *fairness*, *innovation*, *pressure*, *recognition*, *support*, and *trust*. A brief definition/description of each of these dimensions is provided in Table 1.

In the next section, while presenting our research model and associated propositions, we will discuss how each of these dimensions would be expected to moderate the relationship between an individual's perceptions (of an innovation) and behavioral intention (to adopt and use it). Briefly, however, we will take a couple of these climate dimensions (*support* and *autonomy*) and discuss the relevance of these dimensions of organizational climate for the adoption of technological innovations.

Senior management's attitude toward change (consequential to the introduction of technology innovations) and thus the extent of their *sup-*

port impacts the adoption of these technology innovations (Damanpour, 1991). Senior management teams may be very conservative, preferring the status quo and using current or time-tested methods innovating only when they are seriously challenged by their competition or by shifting consumer preferences (Miller & Friesen, 1982). By contrast, they may be risk prone, actually encouraging and actively supporting the use of innovative techniques to move the organization forward, usually trying to obtain a competitive advantage by routinely making dramatic innovative changes and taking the inherent risks associated with those innovations (Litwin & Stringer, 1968). The potentially disruptive features typically associated with the adoption of (radical) innovations require an organizational context where managers encourage individual members of the organization to take (prudent levels of) risk, support adoption of technology innovations, and be supportive of changes in their organizations (Dewar & Dutton, 1986). Organizations should be wary, however, that a follower approach taken by employees

Figure 2. Research model



may promote a “mindless” environment resulting in undesirable levels of risk-taking, which can cause significant problems (Swanson & Ramiller, 2004).

Organizational context/climate also reflects the extent of focus on autonomy/empowerment vs. control of its members. An organic organization as contrasted with mechanistic organization is typically associated with open and free-flowing communication, sharing of necessary information and knowledge, flexibility, and absence of rigid rules and regulations; such an organization context is usually positively related to innovation (Aiken & Hage, 1971; Kimberly & Evanisko, 1981). Furthermore, an organizational climate that is geared toward and has built-in expectation of high levels of achievement and high standards of excellence nurtures a vibrant base of challenges posed to its members who have the freedom to apply innovative technologies, techniques, and procedures to effectively accomplish the tasks (Rosenthal & Crain, 1963). Such an organizational context will be more prone to encouraging its members to adopt technology innovations to accomplish high levels of performance.

RESEARCH MODEL AND TENTATIVE PROPOSITIONS

Based on the foregoing brief discussion of the extant research, we extend the standard TAM model with social context dimensions as shown in Figure 2.

Traditional TAM Propositions

An individual’s intention to adopt/use technology is driven by his or her perceptions of the usefulness of the technology (Davis et al., 1989). This contention, as noted in the background research section, has been supported extensively in previous research (Agarwal & Prasad, 1999; Davis et al., 1989; Hu et al., 1999; Jackson et al., 1997; Venkatesh, 1999, 2000; Venkatesh & Davis, 2000; Venkatesh & Davis, 1996; Venkatesh & Morris, 2000). A primary reason why individuals would intend to adopt/use mobile computing devices for B2B transactions is that they believe that this technology will provide them the flexibility to perform their job and enable their job performance enhancement (Davis, 2002; Intel, 2003). Fur-

thermore, following the findings of Karahanna et al. (1999), the perceived usefulness of an IT influences the attitudes of both potential adopters and users of an IT. However, we contend that when an IT has demonstrated its usefulness over time, it is likely to play a significant role in the overall infusion of the technology. Therefore, we propose:

- **Proposition 1:** Perceived usefulness will have a positive effect on organizational members' intention to adopt/use mobile computing devices for B2B transactions.
- **Proposition 2:** Actual/demonstrated usefulness will have a positive effect on organizational members' continued usage of mobile computing devices for B2B transactions.

As noted earlier, the second major determinant of behavioral intentions in the TAM model, *perceived ease of use*, has been observed to have both a (somewhat weak) direct influence on behavioral intention as well as a (strong) indirect influence through its effect on perceived usefulness (Davis, 1989; Davis et al., 1989; Hu et al., 1999; Jackson et al., 1997). This is understandable since a person who believes that a technology innovation is (relatively) easy to understand and use, and is less demanding of efforts, would likely believe that using such a technology is also more useful. While perceived ease of use may trigger users' intention to adopt/use the innovation (mobile computing devices for B2B), it is unlikely to play a key role in the spread/infusion since users would likely become more familiar with all the features of the innovation and gain significant expertise with time following the initial use. Hence, we propose the two following propositions:

- **Proposition 3a:** Perceived ease of use will have a positive effect on organizational members' intention to adopt/use

mobile computing devices for B2B transactions.

- **Proposition 3b:** Perceived ease of use will positively influence organizational members' perceptions of the usefulness of mobile computing devices for B2B transactions.

Extended TAM Propositions

One of the key objectives of this chapter is to examine what role, if any, social context plays in the link between individuals' perceptions of usefulness/ease of use and behavioral intentions of the TAM model. We pointed out that social context, when conceptualized in terms of climate/culture, is a multi-dimensional construct composed of eight dimensions (Koys & DeCotiis, 1991). Since there has been no attempt to examine this additional set of dimensions within the context of TAM, many of the arguments and much of the rationale that we provide in the rest of this section while developing the propositions are likely to be tentative.

Autonomy

At one end of the spectrum, an organization can be extremely control and compliance oriented (*mechanistic* organizational context) in formulating, administering, and closely monitoring and enforcing a set of policies and procedures that guide employee work activities. At the opposite end of the spectrum, an organization can be performance and achievement oriented (*organic* organizational context) by empowering their employees to determine their task priorities and schedule, providing them the autonomy to make use of any and all techniques, tools, and technologies that they deem best for getting the work done, and being flexible with respect to adherence on the standard policies and procedures. Thus, organizations where the members perceive greater *autonomy* and flex-

ibility being provided to them in making decisions and choices on their task-related activities are likely to more quickly exploit (any) opportunity that technology innovations offer. While this is fairly obvious when the technology is perceived to be useful and easy to use, even in instances where such perceptions (of ease of use and usefulness) may not be completely true, the organizational members may still be more willing to make informed decisions that they are responsible and accountable for (Aiken & Hage, 1971; Kimberly & Evanisko, 1981). To become better informed, they may actively seek out knowledge from various pockets of the (internal) organization as well as from external sources (e.g., consultants, vendors, trade literature, etc.). Therefore:

- **Proposition 4a:** The relationship between employees' perceptions (of usefulness and ease of use of the technology) and their intentions to adopt/use mobile computing devices for B2B transactions will be stronger in organizational contexts that provide greater autonomy to their employees.
- **Proposition 4b:** The relationship between the actual/demonstrated usefulness and continued usage of mobile computing devices for B2B transactions will be stronger in organizational contexts that provide greater autonomy to their employees.

Cohesion

As would be noted from the brief description provided in Table 1, *cohesion* refers to an organizational context/climate that fosters a sense of sharing, caring, accommodation, and togetherness among the members/employees (Koys & DeCotiis, 1991). Communication, sharing, and exchange of information and knowledge amongst the members is bound to be much more open in such a context. Employees would more willingly share their experiences and sup-

port one another when attempting to make decisions on complex and unknown topic areas (e.g., relevance and mastery of new technologies). It is, therefore, reasonable to expect that potential adopters of new technology innovations (mobile computing devices) would be more willing and prepared to assume any challenges posed by the new technology environment in view of the potential support that they can expect from their colleagues in their work environment. Therefore, we propose the following:

Proposition 5a: The relationship between employees' perceptions (of usefulness and ease of use of the technology) and their intentions to adopt/use mobile computing devices for B2B transactions will be stronger in organizational contexts that foster a greater sense of cohesion/cohesiveness among their employees.

Proposition 5b: The relationship between the actual/demonstrated usefulness and continued usage of mobile computing devices for B2B transactions will be stronger in organizational contexts that foster a greater sense of cohesion/cohesiveness among their employees.

Trust

The third dimension of organizational climate, *trust*, refers to the extent to which employees within the organization can openly communicate with their superiors, seek their guidance and expertise, and be confident that the integrity of sensitive information will not be compromised (Koys & DeCotiis, 1991). It is easy to visualize that such expectations of trust work in both directions—from subordinate to superiors and vice versa. Trust also involves an expectation of confidence in the goodwill of others in the organizational context/environment, as well as the prospects for continuity of the relationship entered into (Hart & Saunders, 1997). It is normal to expect that in these trusting organizational contexts, employees will be more pre-

pared to share their difficulties and concerns (work related and even personal), propose potential technology-based solutions, and seek approval/guidance/advice from their superiors and peers. This can be quite important as in the case of introduction of mobile computing devices where the work arrangements and workflows are bound to be disrupted and changed quite radically (e.g., employees may not have to be always present on site and could increasingly work from off-site locations, at home, or on the move). Trust is a significant determinant of a stable relationship (Mayer, Davis, & Schoorman, 1995; McKnight, Choudhury, & Kacmar, 2002). Therefore, we propose:

- **Proposition 6a:** The relationship between employees' perceptions (of usefulness and ease of use of the technology) and their intentions to adopt/use mobile computing devices for B2B transactions will be stronger in organizational contexts that promote and reinforce trust between employees and the organization.
- **Proposition 6b:** The relationship between the actual/demonstrated usefulness and continued usage of mobile computing devices for B2B transactions will be stronger in organizational contexts that promote and reinforce trust between employees and the organization.

Pressure

The fourth dimension of organizational climate, *pressure*, refers to the fact that the work context may not provide adequate time for the employees to accomplish their task-related activities and achieve the required standards of performance and goals (Koys & DeCotiis, 1991). Typically, it would be reflective of a situation of significant stress, perhaps hasty

decisions and actions resulting in suboptimal results, and generally chaos. However, such a stressful environment may also be one that could spur the organizational members to creatively look for (technologically) innovative solutions to alleviate the difficulties and infuse some order. To the extent that the performance of tasks is not geographically constrained (e.g., assembly-line work in automotive manufacturing, patrons being serviced in a restaurant or a bank), it is possible that mobile computing devices may indeed alleviate the time pressure that is so rampant in the work context. For example, employees may become skillful in time management through the convenience of mobile computing devices in coordinating work and personal tasks (Davis, 2002; Intel, 2003). Therefore, surprising and counter-intuitive as it might sound, we propose:

- **Proposition 7a:** The relationship between employees' perceptions (of usefulness and ease of use of the technology) and their intentions to adopt/use mobile computing devices for B2B transactions is likely to be stronger in organizational contexts that reflect one of (time) pressure for employees to accomplish their task and realize the set performance standards.
- **Proposition 7b:** The relationship between the actual/demonstrated usefulness and continued usage of mobile computing devices for B2B transactions is likely to be stronger in organizational contexts that reflect one of (time) pressure for employees to accomplish their task and realize the set performance standards.

Support

The fifth dimension of organizational climate, *support*, reflects an organizational context that is tolerant of errors and mistakes that employ-

ees may commit, and is supportive of them as long as they learn from these (Koys & DeCotiis, 1991). An environment that is permissive and lets its members learn from mistakes without fear of punishment and reprisal could engender deep-rooted learning, a “can-do” attitude to problem solving, and (reasonable) risk-taking orientation (Litwin & Stringer, 1968). As noted earlier, management’s attitude toward change (often triggered by the introduction of technology innovations) and thus the extent of their support impacts the adoption and successful implementation of these technology innovations (Damanpour, 1991; Sanders & Courtney, 1985). The potentially disruptive features typically associated with the adoption of (radical technology) innovations require an organization context where managers encourage individual members of the organization to take (prudent levels of) risk, support adoption of technology innovations, and be supportive of changes in their organizations (Dewar & Dutton, 1986). Supportive organizational context is also conducive to successful IT implementation (Ramamurthy, Premkumar, & Crum, 1999). Caron, Jarvenpaa, and Stoddard (1994) chronicle how CIGNA Corporation, due to its supportive and tolerance-for-failure environment, facilitated significant learning to accrue in the context of major disruptive and radical changes triggered by business process reengineering projects. Therefore, we propose:

- **Proposition 8a:** The relationship between employees’ perceptions (of usefulness and ease of use of the technology) and their intentions to adopt/use mobile computing devices for B2B transactions is likely to be stronger in organizational contexts that are tolerant and supportive of employees in accomplishing their work.
- **Proposition 8b:** The relationship between the actual/demonstrated usefulness and

continued usage of mobile computing devices for B2B transactions is likely to be stronger in organizational contexts that are tolerant and supportive of employees in accomplishing their work.

Recognition

The sixth dimension of organizational climate, *recognition*, reflects an organizational context where employee achievements and accomplishments are acknowledged and recognized (Koys & DeCotiis, 1991). *Human relations management* and *job enrichment* literature (Hackman & Oldham, 1980) points out that intrinsic rewards (e.g., employee-of-the-month recognition) at times are more important than extrinsic rewards (e.g., salary raises, promotion). Extrinsic and intrinsic motivation literature has also been used significantly to explain adoption and use of innovations (Davis, Bagozzi, & Warshaw, 1992). Resource-based theory also acknowledges the vital role human assets/resources play in contemporary hyper-competitive external environments where progressive organizations strive to keep their employees satisfied and thus retain top talent. It is, therefore, natural to expect that organizations should strive to create a climate that spurs their employees to constantly look out for creative solutions (including new technology innovations) that foster excellence in achievement. Obviously, this is unlikely when such efforts and accomplishments go unrecognized. Thus, we would propose:

- **Proposition 9a:** The relationship between employees’ perceptions (of usefulness and ease of use of the technology) and their intentions to adopt/use mobile computing devices for B2B transactions is likely to be stronger in organizational contexts that are open to acknowledge and recognize

- the accomplishments of their employees.
- **Proposition 9b:** The relationship between the actual/demonstrated usefulness and continued usage of mobile computing devices for B2B transactions is likely to be stronger in organizational contexts that are open to acknowledge and recognize the accomplishments of their employees.

Fairness

The seventh dimension of organizational climate, *fairness*, reflects an organizational context where employees believe in equitable and non-arbitrary treatment (Koys & DeCotiis, 1991). This reinforces the notion that hard, sincere, and smart work pays off. Individuals that believe an inequity exists, for example, are likely to resent and resist organizational changes (Joshi, 1989, 1991). Clearly an organization that does not design its workplace context with work/job assignments that are perceived to be fair and rewards that are perceived to be equitable for similar accomplishments would trigger significant discontent and distrust. Such an environment is hardly likely to evoke any voluntary or enthusiastic response to work-related organizational challenges, including searching for new technology innovations. Therefore, we would propose:

- **Proposition 10a:** The relationship between employees' perceptions (of usefulness and ease of use of the technology) and their intentions to adopt/use mobile computing devices for B2B transactions is likely to be stronger in organizational contexts that are deemed to be fair in the treatment of their employees.
- **Proposition 10b:** The relationship between the actual/demonstrated usefulness and continued usage of mobile computing devices for B2B transactions is likely to be stronger in organizational contexts that

are deemed to be fair in the treatment of their employees.

Innovation

The last (eighth) dimension of organizational climate, *innovation*, reflects an organizational context where employees believe change from status-quo can be good, that originality is valued, and risk taking will be encouraged (Koys & DeCotiis, 1991). As noted earlier, management's attitude toward change (often triggered by the introduction of technology innovations) impacts the adoption of these technology innovations (Damanpour, 1991). Some senior management teams may have conservative attitudes toward innovation and associated risk, preferring the status quo and using current or time-tested methods; such organizations innovate only when they are seriously challenged by their competition or by shifting consumer preferences (Miller & Friesen, 1982). By contrast, other senior management teams may be risk prone, actually encouraging and actively supporting the use of innovative techniques to move the organization forward. Such organizations usually try to obtain a competitive advantage by routinely making dramatic innovative changes and taking the inherent risks associated with those innovations. The potentially disruptive features typically associated with the adoption of (radical technology) innovations require an organization context where managers encourage individual members of the organization to take prudent levels of risk, support adoption of technology innovations, and be supportive of changes in their organizations (Dewar & Dutton, 1986). Thus, we would propose:

- **Proposition 11a:** The relationship between employees' perceptions (of usefulness and ease of use of the technology) and their intentions to adopt/use mobile computing devices for B2B transactions

is likely to be stronger in progressive/innovative organizational contexts.

- **Proposition 11b:** The relationship between the actual/demonstrated usefulness and continued usage of mobile computing devices for B2B transactions is likely to be stronger in progressive/innovative organizational contexts.

B2B APPLICATION DOMAIN AND SUGGESTED RESEARCH METHODOLOGY

Some of the broad domains of B2B application areas that are relevant for mobile-computing and of interest to us for this research would be inventory management, customer relationship and service management, sales force automation, product locating and purchasing, dispatching and diagnosis support to, say, technicians in remote locations, mobile shop-floor quality control systems, as well as those applications and transactions in supply chain management (SCM) that facilitate the integration of business processes along the supply chain (Rao & Minakakis, 2003; Turban, King, Lee, & Viehland, 2004; Varshney & Vetter, 2001). An example of B2B transactions in the SCM context includes data transmission from one business partner to another through the typical enterprise resource planning (ERP) interactions. Other scenarios may involve the ability to continue working on projects while in transit or the ubiquitous access to documents via “hot spots” or wireless network access (Intel, 2003). Consequently, in light of the fact that a number of application domains have preexisted the Internet, the choice of application areas could be either Internet or non-Internet based.

As noted before, mobile computing is still in a very early stage of its evolution and use within organizations in a B2B context. Although a

large-scale field survey would be required to test the research model that we presented, such an approach may not be appropriate in this context due to the exploratory nature of the inquiry proposed here. Therefore, the research methodology that we suggest and propose that researchers use at this stage is a combination of both qualitative and quantitative research for data collection. Rather than a large national random sample, we propose a purposive convenience-based sample of a few (say, 8-12) large and medium-sized corporations with almost equal composition of manufacturing and service sectors. Furthermore, based on secondary information and personal contacts, we would prefer that researchers select an equal mix of corporations that do not (yet) use and those that currently use mobile computing so that we can capture their “intention” and subsequently their “continued use.” Although the “social context” or “climate” prevailing within each of these organizations may be a “given reality” at least at a point in time, as observed in most past research, it is the interpretations of this social context/climate that would drive individual actions, especially when the intended/actual behavior (in this case, adoption and use of mobile computing) is not mandatory (Moran & Volkwein, 1992). Thus, in-depth interviews coupled with a questionnaire survey from a number of focal members (about 20 to 25), sampled from multiple functional areas (that are amenable for use of mobile computing devices such as sales and marketing, purchasing, and operations) within participant organizations, should be used to capture individual perceptions of the mobile computing devices and their organization’s social context. As argued above, since the rate of diffusion for mobile computing devices for B2B transactions is still relatively small, a convenient sampling approach among organizations that have and have yet to adopt these technologies is appro-

priate. To ensure relevance and reasonable generalizability of the study findings of the convenience-based sampling suggested by us, participants from each organization should be chosen randomly. A number of statistical techniques such as logistic regression (for the “intention to adopt” stage) and structural equation modeling or hierarchical moderated regression analysis (for the “infusion” stage) would be candidates for data analyses.

CONCLUSION

In this chapter we incorporated the social context of an organization into TAM and proposed an extended model to investigate adoption/use of mobile computing devices for B2B transactions as a technological innovation. We believe that such an extension is appropriate because aspects of social context have in general been found significant with the introduction of new technologies. In particular, a micro-level analysis of this phenomenon for the introduction of new technologies is rare. Since the unit of analysis of this chapter is individual employees, we utilized dimensions of psychological climate to represent the social context of an organization. The primary objective of this chapter was to posit how perceptions of the social context of an organization would moderate the intention to adopt/use and infusion of a technology innovation.

A key feature of this study is that we examined an information technology that has the potential of becoming a dominant paradigm and platform for future computing applications. As we noted, although mobile computing devices have existed for several years, their use for business-to-business transactions or operating context has not been adequately or systematically explored in academic research. We drew upon theories from the diffusion of inno-

vation, information systems, and organizational behavior literature, among others, to develop our research model and the associated 10 propositions. The model we proposed could serve as a foundation for one stream of IS research that integrates social context of an organization into TAM to examine the vital role of mobile computing devices in electronic commerce.

IMPLICATIONS, LIMITATIONS, AND FUTURE RESEARCH DIRECTIONS

Since the empirical segment of this research has not yet been conducted, we can only conjecture several potential research contributions for researchers and practitioners. One implication that this work has for future research is the exploration of how the social context of an organization may influence the acceptance and spread of an information technology innovation. The social context of an organization has not been applied to TAM, and an extension focusing on the micro-level aspects of the social context have not been widely examined in the literature. By explicitly investigating the social context of an organization, this study extends the innovation adoption and TAM literature base. Our model may be considered a viable and prudent alternative to the UTAUT model. Utilizing a (valid and popular base) model and measures that have become widely accepted among IS researchers allows for researchers in future research to replicate our study and examine other factors of interest. This chapter also addresses the need to explore technology that is close to the “leading edge” (Lyytinen, 1999, p. 26), which is recommended for maintaining the relevance of IS research (Benbasat & Zmud, 1999; Lyytinen, 1999; Orlikowski & Iacono, 2001). Obviously, considerable care and precautions (in the design of

the study, operationalization, and evaluation of the measurement properties) will be needed in translating the theoretical model proposed in this chapter into a large-scale empirical investigation that can establish validity and reliability of its results.

The potential implication that this work has for IS practice is that it identifies a number of contextual factors that may influence the acceptance of a technological innovation that an organization wishes to introduce. Mobile computing devices can enhance employee productivity by granting them flexibility in work location and time management (Intel, 2003). Organizations that covet such gains in productivity are likely cognizant of the investments typically at stake when implementing IT innovations. Given that aspects of the social context of an organization are suggested as having a significant role in the introduction of mobile computing devices (Jessup & Robey, 2002; Sarker & Wells, 2003), it is desirable to understand the influence that the social context of an organization plays. Moran and Volkwein (1992) state that focusing on the micro-level aspect of the social context is appealing because it is relatively accessible, more malleable, and the appropriate level to target short-term interventions aimed at producing positive organizational change. This study helps to uncover several future opportunities for organizations since mobile computing devices have the potential to become the dominant paradigm for future computing applications (March et al., 2000; Sarker & Wells, 2003).

Although this chapter offers several potential contributions, several limitations exist. The social context of an organization is operationalized through psychological climate dimensions. The definition of social context that we adopted takes a much broader view than focusing on the individual incorporating traditions from research in the organizational culture literature as well. We feel that it is

appropriate to use the social context of the organization to begin the integration of culture and climate literature. It is our opinion that the psychological climate research is the most appropriate theory to support the research model, which presents opportunities in future work to examine other aspects of the social context of an organization that may be influential in the acceptance of a technological innovation. Another limitation of this study (when an empirical investigation is conducted) is that it may obtain retrospective accounts/information from (current) users of mobile computing devices. Retrospective accounts are an issue because individuals may not be able to accurately recall the past. It would be necessary to consider preventive measures on this front to ensure validity and reliability of the results.

REFERENCES

- Agarwal, R., & Prasad, J. (1999). Are individual differences germane to the acceptance of new information technologies? *Decision Sciences*, 30(2), 361-391.
- Aiken, M., & Hage, J. (1971). The organic organization and innovation. *Sociology*, 5, 63-82.
- Astley, W., & Van de Ven, A. (1983). Central perspectives and debates in organizational theory. *Administrative Science Quarterly*, 28, 245-273.
- Benbasat, I., & Zmud, R. W. (1999). Empirical research in information systems: The practice of relevance. *MIS Quarterly*, 23(1), 3-16.
- Boudreau, M., Loch, K., Robey, D., & Straub, D. (1998). Going global: Using information technology to advance the competitiveness of the virtual transnational organization. *Academy of Management Executive*, 12(4), 120-128.

- Caron, J., Jarvenpaa, S., & Stoddard, D. (1994). Business reengineering at CIGNA Corporation: Experiences and lessons learned from the first five years. *MIS Quarterly*, 18(3), 233-250.
- Damanpour, F. (1991). Organizational innovation: A meta-analysis of effects of determinants and moderators. *Academy of Management Journal*, 34, 555-590.
- Davis, F. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3), 319-340.
- Davis, F., Bagozzi, R., & Warshaw, P. (1989). User acceptance of computer technology: A comparison of two theoretical models. *Management Science*, 35(8), 982-1003.
- Davis, F., Bagozzi, R., & Warshaw, P. (1992). Extrinsic and intrinsic motivation to use computers in the workplace. *Journal of Applied Social Psychology*, 22, 1111-1132.
- Davis, G. (2002). Anytime/anyplace computing and the future of knowledge work. *Communications of the ACM*, 45(12), 67-73.
- Denison, D. (1996). What is the difference between organizational culture and organizational climate? *Academy of Management Review*, 21(3), 619-654.
- Denison, D., & Mishra, A. (1995). Toward a theory of organizational culture and effectiveness. *Organization Science*, 6(2), 204-223.
- Dewar, R., & Dutton, J. (1986). The adoption of radical and incremental innovation: An empirical analysis. *Management Science*, 23, 1422-1433.
- Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention, and behavior: An introduction to theory and research*. Reading, MA: Addison-Wesley.
- Glick, W. (1985). Conceptualizing and measuring organizational and psychological climate: Pitfalls in multilevel research. *Academy of Management Review*, 10(3), 601-616.
- Hackman, J., & Oldham, G. (1980). *Work redesign*. Reading, MA: Addison-Wesley.
- Harrison, A., & Rainer, R. (1992). The influence of individual differences on skill in end-user computing. *Journal of Management Information Systems*, 9(1), 93-111.
- Hart, P., & Saunders, C. (1997). Power and trust: Critical factors in the adoption and use of electronic data interchange. *Organization Science*, 8(1), 23-42.
- Hartwick, J., & Barki, H. (1994). Explaining the role of user participation in information system use. *Management Science*, 40(4), 440-465.
- Hu, P., Chau, P., Sheng, O., & Tam, K. (1999). Examining the Technology Acceptance Model using physician acceptance of telemedicine technology. *Journal of Management Information Systems*, 16(2), 91-112.
- Intel. (2003). *Effects of wireless mobile technology on employee productivity*. Intel Information Technology White Paper (pp. 1-20), USA.
- Jackson, C., Chow, S., & Leitch, R. (1997). Toward an understanding of the behavioral intention to use an information system. *Decision Sciences*, 28(2), 357-389.
- Jessup, L., & Robey, D. (2002). The relevance of social issues in ubiquitous computing environments. *Communications of the ACM*, 45(12), 88-91.
- Joshi, K. (1989). The measurement of fairness or equity perceptions of management information systems users. *MIS Quarterly*, 13(3), 343-358.

- Joshi, K. (1991). A model of users' perspective on change: The case of information systems technology implementation. *MIS Quarterly*, 15(2), 229-242.
- Karahanna, E., Straub, D., & Chervany, N. (1999). Information technology adoption across time: A cross-sectional comparison of pre-adoption and post-adoption beliefs. *MIS Quarterly*, 23(2), 183-213.
- Kimberly, J., & Evanisko, M. (1981). Organizational innovation: The influence of individual, organizational and contextual factors on hospital adoption of technological and administrative innovations. *Academy of Management Journal*, 24, 689-713.
- Koys, D., & DeCotiis, T. (1991). Inductive measures of psychological climate. *Human Relations*, 44(3), 265-283.
- Legler, R., & Reischl, T. (2003). The relationship of key factors in the process of collaboration. *The Journal of Applied Behavioral Science*, 39(1), 53-72.
- Litwin, G., & Stringer, R. (1968). *Motivation and organizational climate*. Boston: Harvard University Press.
- Lyytinen, K. (1999). Empirical research in information systems: On the relevance of practice in thinking of IS research. *MIS Quarterly*, 23(1), 25-28.
- Lyytinen, K., & Yoo, Y. (2002a). Issues and challenges in ubiquitous computing. *Communications of the ACM*, 45(12), 63-65.
- Lyytinen, K., & Yoo, Y. (2002b). Research commentary: The next wave of nomadic computing. *Information Systems Research*, 13(4), 377-388.
- March, S., Hevner, A., & Ram, S. (2000). Research commentary: An agenda for information technology research in heterogeneous and distributed environments. *Information Systems Research*, 11(4), 327-341.
- Mathieson, K. (1991). Predicting user intentions: Comparing the Technology Acceptance Model with the Theory of Planned Behavior. *Information Systems Research*, 2(3), 173-191.
- Mayer, R., Davis, J., & Schoorman, F. (1995). An integrative model of organizational trust. *Academy of Management Review*, 20(3), 709-734.
- McKnight, D., Choudhury, V., & Kacmar, C. (2002). Developing and validating trust measures for e-commerce: An integrative typology. *Information Systems Research*, 13(3), 334-359.
- Miller, D., & Friesen, P. (1982). Innovation in conservative and entrepreneurial firms: Two modes of strategic momentum. *Strategic Management Journal*, 3, 1-25.
- Moore, G., & Benbasat, I. (1991). Development of an instrument to measure the perceptions of adopting an information technology innovation. *Information Systems Research*, 2(3), 192-222.
- Moran, E., & Volkwein, J. (1992). The cultural approach to the formation of organizational climate. *Human Relations*, 45, 19-47.
- Nelson, D. (1990). Individual adjustment to information-driven technologies: A critical review. *MIS Quarterly*, 14(1), 79-98.
- Orlikowski, W. (1993). Learning from notes: Organizational issues in groupware implementation. *The Information Society*, 9, 223-250.
- Orlikowski, W., & Iacono, C. (2001). Research commentary: Desperately seeking the "IT" in IT research—A call to theorizing the IT artifact. *Information Systems Research*, 12(2), 121-134.

- Pareek, U. (1987). *Motivating organizational roles*. New Delhi, India: Oxford and IBH.
- Pfeffer, J. (1982). *Organizations and organizational theory*. Boston: Pitman.
- Ramamurthy, K., Premkumar, G., & Crum, M. (1999). Organizational and inter-organizational determinants of the EDI diffusion: A causal model. *Journal of Organizational Computing and Electronic Commerce*, 9(4), 253-285.
- Rao, B., & Minakakis, L. (2003). Evolution of mobile location-based services. *Communications of the ACM*, 46(12), 61-65.
- Rogers, E. (2003). *Diffusion of innovations* (5th ed.). New York: The Free Press.
- Rosenthal, D., & Crain, R. (1963). Executive leadership and community innovation: The fluoridation experience. *Urban Affairs Quarterly*, 1, 39-57.
- Sanders, L., & Courtney, J. (1985). A field study of organizational factors influencing DSS success. *MIS Quarterly*, 9(1), 77-93.
- Sarker, S., & Wells, J. (2003). Understanding mobile handheld device use and adoption. *Communications of the ACM*, 46(12), 35-40.
- Satyanarayanan, M. (1996). Fundamental challenges in mobile computing. *Proceedings of the ACM Symposium—Principles of Distributed Computing*, Philadelphia.
- Schneider, B. (Ed.). (1990). *Organizational climate and culture*. San Francisco: Jossey-Bass.
- Straub, D. (1994). The effect of culture on IT diffusion: E-mail & fax in Japan and the U.S. *Information Systems Research*, 5(1), 23-47.
- Swanson, E. B., & Ramiller, N. C. (2004). Innovating mindfully with information technology. *MIS Quarterly*, 28(4), 553-583.
- Taylor, S., & Todd, P. (1995). Understanding information technology usage: A test of competing models. *Information Systems Research*, 6(2), 144-176.
- Turban, E., King, D., Lee, J., & Viehland, D. (2004). *Electronic commerce: A managerial perspective*. Upper Saddle River, NJ: Prentice-Hall.
- Van de Ven, A. (1986). Central problems in the management of innovation. *Management Science*, 32, 590-607.
- Varshney, U., & Vetter, R. (2001). A framework for the emerging m-commerce applications. *Proceedings of the 34th Hawaii International Conference on Systems Sciences*.
- Venkatesh, V. (1999). Creation of favorable user perceptions: Exploring the role of intrinsic motivation. *MIS Quarterly*, 23(2), 239-260.
- Venkatesh, V. (2000). Determinants of perceived ease of use: Integrating control, intrinsic motivation, and emotion into the Technology Acceptance Model. *Information Systems Research*, 11(4), 342-365.
- Venkatesh, V., & Davis, F. (2000). A theoretical extension of the Technology Acceptance Model: Four longitudinal field studies. *Management Science*, 46(2), 186-204.
- Venkatesh, V., & Davis, F. D. (1996). A model of the antecedents of perceived ease of use: development and test. *Decision Sciences*, 27(3), 451-481.
- Venkatesh, V., & Morris, M. (2000). Why don't men ever stop to ask for directions? Gender, social influence, and their role in technology acceptance and usage behavior. *MIS Quarterly*, 24(1), 115-139.
- Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of infor-

mation technology: Toward a unified view. *MIS Quarterly*, 27(3), 425-478.

Zammuto, R., & O'Connor, E. (1992). Gaining advanced manufacturing technologies' benefits: The roles of organization design and culture. *Academy of Management Review*, 17(4), 701-728.

Zmud, R. (1979). Individual differences and MIS success: A review of the empirical literature. *Management Science*, 25(10), 966-979.

Zmud, R. (1982). Diffusion of modern software practices: Influence of centralization and formalization. *Management Science*, 28, 1421-1431.

ENDNOTE

- ¹ We use the term *infusion* to refer to diffusion and spread of the innovation within an organization's internal environment.

Chapter XLIX

A Socio–Cultural Analysis of the Present and the Future of the M–Commerce Industry

Ritanjan Das

University of Portsmouth, UK

Jia Jia Wang

University of Bradford, UK

Pouwan Lei

University of Bradford, UK

ABSTRACT

With high optimism, the third generation mobile communication technologies were launched and adopted by telecommunication giants in different parts of the globe—Hutchison 3G in the UK, Verizon in the USA and NTT DoCoMo in Japan. However, with an uncertain and turbulent social, economic and political environment, and the downturn in the global economy, difficult conditions are pronounced for the initial promises of m-commerce technologies to be fully realized. The causes for this, determined so far, have been largely of a technical nature. In this chapter, we shift the focus of analysis from a pure technical approach to a socio-cultural one. The basic premise of the chapter is that cultural variations do play a very important part in shaping potential consumers' choice, belief and attitude about m-commerce services. We believe that to be an important way for the m-commerce industry to fulfill its potential.

INTRODUCTION

This chapter discusses the impact of socio-cultural aspects on mobile commerce (m-commerce). While m-commerce heralds the next revolutionary phase in the advent of digital technology, still the digital industry can be con-

sidered in its infancy. This makes its specific categorization difficult. However, as Mahatanankoon, Wen, and Lim (2004) point out, the 1980s can be roughly classified as the age of PCs, the '90s as the “decade of the Internet, and...the first decade of the 21st century as the decade of mobile computing and

mobile commerce.” By the end of 2004, the number of mobile phone subscribers was expected to be 1.5 billion—about one-quarter of the world’s population (Evans, 2004). The ITU (International Telecommunication Union) said the growth in mobile phone subscribers outpaced growth in the number of users of fixed lines (1.185 billion) today and is outstripping the rate of increase in Internet users. Emerging markets such as China, India, and Russia contribute to the growth. The current state of the digital age that we live in convinces us of the remarkable rate that wireless data communication (WDC)/mobile computing/mobile commerce services are penetrating the market with. It is predicted to be one of the main driving forces for the computing industry, as well as a substantial revenue-generating platform for businesses. Recent major findings by the research firm IDC (Mahatanankoon et al., 2004) predicted a growth in the mobile commerce revenues from US\$500 million in 2002 to US\$27 billion by 2005. Predictions by Forrester Research (Mahatanankoon et al., 2004) estimate an average of 2.2 wireless phones per U.S. household by 2007, with up to 2.3 million wired phone subscribers making a switch to wireless services. Worldwide, there were 94.9 million users of m-commerce in 2003; this is expected to grow to 1.67 billion in 2008, resulting in estimated global revenue of US\$554.37 billion (Wireless Week, 2004).

UNDERSTANDING M-COMMERCE

Mobile-commerce can be defined as the commercial transactions conducted through a variety of mobile equipment over a wireless telecommunication network in a wireless environment (Barnes, 2002; Coursaris & Hassanein, 2002; Gunsaekaran & Ngai, 2003). Currently these wireless devices include two-way pagers/SMS (short message systems), WAP-

(wireless application protocol) equipped mobile phones, PDAs (personal digital assistants), Internet-enabled laptop computers with wireless access capacity, and consumer premise IEEE 802.11 (a/b) wireless network devices (Leung & Antypas, 2001). The range of applications that characterize m-commerce activities can be largely divided into:

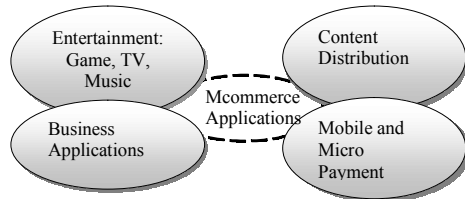
- **Entertainment:** Includes online TV broadcasts, online mobile games, and downloaded music or ring tones.
- **Content Delivery:** Includes reporting, notification, consultation, and so forth.
- **Transactions:** Includes data entry, purchasing, promotions, and so forth (Balasubramanian, Peterson, & Jarvenpaa, 2002; Leung & Antypas, 2001).

Wireless cellular technology (third- and fourth-generation wireless cellular networks) areas have witnessed exciting innovations in recent years. 3G cellular networks offer broadband transmission with speeds up to 2Mbps, allowing for high-speed wireless access to the Internet, e-commerce transactions, and other information services from any location across the globe. Shim and Shim (2003) describe the not-so-far future of the industry as

...a true wireless broadband cellular system (4G), which can support a much higher bandwidth, global mobility, and tight network security; all at a lower cost. 4G systems should be able to offer a peak speed of more than 100 Mbits per second in stationary mode and an average of 20Mbits per second when in motion. The deployment of 4G technologies will allow the dream of a unified wireless Internet to become a reality.

On the other hand, Wi-Fi (wireless fidelity), wireless area local networks that allow users to surf the Internet while moving, are proliferating

Figure 1. M-commerce applications



at astonishing speed on a global scale. World-wide retail chains like Starbucks and McDonalds offer wireless Internet access to their customers. Wi-Fi offers a fast and stable connection; the data rate is several times faster than 3G. Wi-Fi is an important, new, and disruptive technology to mobile telephone technology, and it may be a watershed for all other m-commerce investment by telecom and content providers in the world of the mobile Internet (Lamont, 2001). In making use of this technology, a mobile phone manufacturer (Nokia) and wireless network manufacturer (Cisco) have been working together closely to produce the Wi-Fi phone. A U.S. mobile telecommunication operator has integrated a network of Wi-Fi hotspots with the existing mobile network systems. In such a way, the tariff of accessing the mobile Internet will be reduced to a budget price. More handheld device users will surf the Internet through their compact mobile phones or PDAs when on the move. Furthermore, WiMax (Worldwide Interoperability for Microwave Access), a low-cost wireless broadband connection in wide area network (WAN), will be rolled out (Cherry, 2004). As the wireless cellular and wireless technologies are converging, the tariff of mobile Internet will be affordable in the future. This rapid development of m-commerce technologies, which was earlier considered to be a mere extension of e-commerce activities, has opened up hitherto unseen business opportunities. It has increased an organization's ability to reach its customers regardless of location and

distance, and has also been successful to a certain extent in creating a consumer demand for more advanced mobile devices with interactive features. While e-commerce is characterized by e-marketplaces, an explosion in m-commerce applications has presented the business world with a fresh set of strategy based on personalized and location-based services. Many new business models have been established around the use of mobile devices which typically have the characteristics of portability, low cost, more personalization, GPS (global positioning system), voice, and so on. These new business models include micro-payment and mobile payment, business services, entertainment, and content distribution services (see Figure 1). Because of their existing customer base, technical expertise, and familiarity with billing, mobile telephone operators are the natural candidates for the provision of mobile and micro-payment services, the latter involving small purchases such as vending and other items.

The mobile phone has become a new personal entertainment medium. A wide range of entertainment services are available on it. These include playing online games, downloading ring tones, watching football video clips, broadcasting live TV, downloading music, and so on. Unsurprisingly, adult mobile services and mobile gambling services are among the fast growing services. According to Juniper research, the total revenue from adult mobile services and mobile gambling services could be worth US\$1 billion and US\$15 billion respectively by 2008 (Kowk, 2004). Law regulators have to stay ahead of the fast-growing development. Content distribution is concerned with providing real-time information, notification, and positioning systems for personalized information by location and advertising. Real-time information such as news, traffic reports, stock prices, and even weather forecasts can be distributed to

mobile phones via the Internet. The information can also be personalized to a user's interests thus achieving a greater degree of personalization and localized services. For example the user's profile such as past behaviour, situation, and location can all be used to determine the relevant information to be supplied at any given time. This in turn can lead to such services being effectively provided through a mobile portal. The mobile network operators (MNOs) can prove to be the central player in providing such services, as they have a number of advantages over other portal players (Tsalgatidou & Veijalainen, 2000). Firstly, they have an existing customer relationship that will lead them, with sufficient ease, to identify the location of the subscriber. Secondly, they already have an existing billing relationship with the customers, while the traditional portals do not. These MNOs can also act as a trusted third party and hence play a dominant role in m-commerce applications. Business applications can be (and to some extent, have been) greatly influenced by these m-commerce applications, especially for companies with remote staff. Extending the existing enterprise resource planning (ERP) systems with mobile functionality will provide remote staff such as sales personnel with real-time corporate and management data. Time and location constraints are reduced and the ability of mobile employees is enhanced. The paperless office becomes a reality. M-commerce offers tremendous potential for businesses to respond quickly in supply chains.

CHARACTERISTICS OF M-COMMERCE

With increasing Internet users across the globe, the m-commerce industry definitely holds great promises for the future. However, to realize the complete potential of this industry, it is ex-

tremely important to understand the unique set of features that guide its development, apart from the general advantages of using e-commerce—that is, efficiency, convenience, broader selections, competitive pricing, rich information, diversity, and so on (Wu & Wang, 2004). While all of these apply to m-commerce applications as well, there is the additional requirement of such services blending into the *world of mobility* by focusing on the moments users will spend on the wireless Internet. Distinctive characteristics of the mobile environment make m-entertainment qualitatively different from fixed-line online entertainment (Baldi & Thaug, 2002). These characteristics are as follows:

- **Accessibility:** The services are available irrespective of location and time. As Mizukoshi, Okino, and Tardy (2001) observe, users of m-commerce services have unique usage patterns and niche access timings in contrast to PC-based Internet users.
- **Ubiquity:** It is possible for the users to enjoy m-commerce services whenever they might feel the need of it. The ubiquity of these services also increases the anonymity of the user in comparison to the wired Internet. Uses of services like gambling and adult entertainment, which are to a large extent socially unexpected, also get promoted as a result of this ubiquitousness.
- **Localization:** The entertainment services can be customized to the user's location, thus allowing for better targeted information and transaction-based entertainment services (e.g., mobile coupons).
- **Reachability:** With the user's permission, it is possible to reach him/her anywhere and anytime, thus allowing for immediate interactions in communication ap-

plications. From a service provider's point of view, this enables the transmission of time-critical information (stock prices, game results, etc.) to the interested clients.

- **Personalization:** As wireless devices are a very personal item, it is possible to customize the interface as per its user's preferences, thus creating an individual relationship with the customer, which in turn will encourage transactions and will also prevent the user from switching between service providers (Granger & Huggins, 2000).

PROMISE VS. REALITY

In spite of such promises and expectations, the situation in which the m-commerce industry finds itself today does not present much scope for such high optimism. As Jarvenpaa, Lang, Takeda, and Tuunainen (2003) write:

The advent of m-commerce has fuelled much anticipation of future possibilities. However, predictions that Internet technologies and wireless communication would greatly benefit both firms and individuals have now come under increasing scrutiny. Uncertain technology standards, the complexities of interactive multimedia applications, and the threat of governmental regulation have all contributed to a deflated vision of mobile commerce.

With tremendous expectations of 3G services, mobile phone operators spent billions of dollars in obtaining 3G licenses, marking their first successful launch in Japan in October 2001. But since then, the global economy has taken an unexpected turn towards the worse. In addition, the interest in 3G being pioneered

by the mobile phone operators has confronted unforeseen low ebb in customer response. In effect, the much anticipated global launch of 3G services has been postponed. There are also inherent criticisms of the 3G mechanisms, especially on the security front.

As far as government policies go, Japan, Korea, a number of European nations, and the United States were moving efficiently towards establishing an m-commerce market. But so far, Japan and Korea are the only markets where the wireless Internet has exceeded expectations, while the European markets did not live up to the hype (Baldi & Thaung, 2002). The U.S. has also found itself not so successful in the race to establish a "functional and interoperable infrastructure for m-commerce" (Shim & Shim, 2003).

The portrayed picture in the early days of m-commerce also finds itself in stark contrast with the reality today, amidst a time of turbulent and uncertain political environment, especially in the post-9/11 world. The financial industry's dependence on telecommunications is well known, and 9/11 provided a clear demonstration of how disruptions to the nation's critical infrastructure can and will close markets and disrupt payment flow (Ferguson, 2002). Hence the adoption of m-commerce applications has definitely suffered a setback after a series of disastrous and malicious events, epitomising themselves in the 9/11 and Enron incidents. The telecommunication sector is to be among the worst-affected ones in terms of job loss and stock prices. Many telecommunication companies—Global Crossing, WorldCom, AT&T, and Level 3 Communications—are already gone or are struggling through bankruptcy (Borland, 2002). The downturn in the global economy has indeed spelt a difficult situation for further developments in the digital world to be widely and freely accepted.

However, although cautiously, the world economy is definitely gaining momentum again; concurrently, mobile technologies have advanced in both hardware and software, resulting in affordable handsets with reasonable cost of 3G mobile services. Furthermore, there is growing interest in 3G from the existing mobile phone users, especially in Asia. While in Korea, the number of 3G users has surpassed 1.7 million; in Japan it grew from 150,000 in January 2003 to 2 million in January 2004 (3GNewsroom.com, 2003; NTT DoCoMo, 2004). In Europe however, the growth is slow. There are only 361,000 3G users in the UK and 453,000 in Italy, despite a huge price lowering campaign (Hutchison Whampoa, 2004). The growth rate in other countries is not promising either.

A CRITICAL DIMENSION OF M-COMMERCE: CULTURAL VARIATIONS

Given this varied distribution of interest towards m-commerce applications, it is perhaps justifiable to shift the focus of analysis of the situation from an application-based rigorous technical discussion to a more soft but relevant socio-cultural one. Several researchers have pointed out that the perceptions and usage patterns of m-commerce differ from one customer segment to the another in terms of age, gender, professional orientation, economic standing, and most importantly, their cultural backgrounds. Consumers from different cultural backgrounds considerably vary in their perceptions, outlooks, and beliefs (Hofstede, 1980), which undoubtedly have a great influence on their choices of mobile technologies/applications. Hence *culture* becomes a highly relevant and critical dimension behind the successful realisation of the initial predictions regarding the m-commerce industry. As Shim

and Shim (2003) write, we need to understand, appreciate, and “leverage subtle but important cultural differences exhibited by individuals.”

Such cultural differences exist across societies, geographical locations, age groups, and so forth. It is beyond the scope of the current study to present a complete description and analysis of all the factors that such variations originate from, which is a large and fascinating field of study in itself. However, to show how such variations influence the diffusion of IT-based communication in general (Straub, 1994) and acceptance of mobile entertainment applications in particular, it will be suitable to draw the reader’s attention to a set of factors identified by Baldi and Thaung (2002) which they see as the determining factors behind the relative success of m-commerce in Japan and other eastern societies as compared to the western societies.

- **Attitude towards Time (Killing Time vs. Saving Time):** People from western societies, especially aged ones, would rather prefer to wait for mature services. As Simon, Brilliant, and Macmillan (2000) note, in a society where saving time is a highly relevant part of daily life, the users must first be shown the value, how mobile entertainment applications “kill time”.
- **Commuting Habits (Public Transport vs. Car):** While the public transport system is a more common way for daily commuters in the east, the phones represent an amusing way to spend the time on the way to work or school. On the other hand Europeans/Americans use less public transportation, but use their cars instead. And it is not possible to use interactive entertainment services while driving.
- **Role of the Mobile Phone (Life Style Component vs. Tool):** The mobile phone is fast becoming an important part of the

eastern society, making a fashion statement (using different handsets for different purposes), adding a lifestyle component, and almost becoming a part of the owner's identity. As a result, using a phones for different entertainment purposes is a common practice in the eastern societies. But the scene is not similar in the western world, where people perceive a mobile phone not a fashion statement, but merely a communication tool.

The basic premise of the world of mobile technology is that it would present its users with their own individual space so that they can enjoy a large degree of freedom, but at the same time increase their connectedness with both the physical and cyber world. But as Jarvenpaa et al. (2003) point out, this increased sense of freedom and connectedness at the same time might be a potential harbour for social disorder:

Excessive mobile use encourages superficiality, indifferent behaviour toward one's surrounding, the privatisation of life-style, and increased opportunity of control of others' lives. Similarly, independence can lead to abuses of freedom, compulsive and self-destructive behaviour, isolation and depression. (Jarvenpaa et al., 2003)

What degree this conflict between interconnectedness and individual freedom might reach once again depends on the particular user's social surrounding, cultural background, and even historical context. While certain cultures naturally emphasise freedom, individual needs, values, and goals, others hold social connectivity as central to their lives. However, although a study of a group's cultural dimensions over individual choices (and vice versa) opens fantastic possibilities, we will have to

restrict ourselves from steering our topic to such a direction, as that lies outside the purview of the topic under discussion. We would conclude this section with recognising individual cultural elements as one of the most (if not the most) important factors influencing people's attitude towards mobile applications/entertainment, and this recognition will be the main basis of the course that our argument will take in the next few sections.

CONSUMER GROUPS AND MARKET STRUCTURES

The variety of consumer groups is another important element that influences the diffusion of m-commerce applications in the world. Although highly related to the cultural issues as discussed in the previous section, the different consumer groups deserve a special mention. The importance behind recognising different groups of consumers lies in the fact that as far as mobile entertainment is concerned, each such group has its own unique set of preferences. These preferences differ according to the varying social environments that each consumer group largely finds itself situated in. Baldi and Thaung (2002) distinguishes the following three major customer groups:

- **Kids and Teenagers:** Are the most prolific users of both PC-based wired Internet as well as the wireless environment. Many kids across different societies are regular and efficient users of mobile devices, and are familiar with services such as games and ring tone downloads. These kids probably will also be the first to link their consoles to wireless phones to play games with their peers, given the amount of free time they enjoy and the importance of socializing in a kid's/teenager's life (Hall,

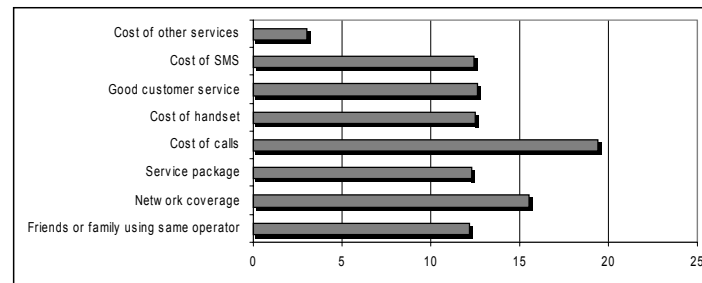
2001). Hence, undoubtedly they will be the most regular, dedicated, and efficient users of the different services mobile technologies can offer.

- **Young Adults:** Especially high school and university students, who have minimal expenses but strong demand for brands, are highly likely to use their mobile phones for daily entertainment services. This group also includes employed people in their mid-to late twenties, who tend to spend a considerable amount on their lifestyle. Familiarity with the digital technology wave, perceptive to new trends, and high rate of education make them a highly likely group to increase the demand for wireless data usage.
- **Business Users:** Although they are unlikely to have a considerable amount of time to gain familiarity with the latest technologies and entertainment services, will use their phones to a large extent for infotainment and communication. Real-time news applications—such as stock market information, share prices, political incidents, weather forecasts, and so forth—will be frequently used and popular services among this group.
The overall market structure is another important determining factor of the mobile industry. Once again, the western and eastern worlds show a number of differences in the way their markets operate. Baldi and Thaung (2002) have also enumerated these as follows:
- **Originator for Phone Specifications (Carrier vs. Manufacturer):** Compatibility amongst different handsets is a problem in European and American markets, as each manufacturer specifies the phone attributes themselves, and there is no central guideline. But the scenario is exactly the opposite in the eastern world, where central regulations exist regarding hand-

set features. For example, in Japan, NTT DoCoMo gives and monitors handset specifications to each of the four major handset manufacturers of the country.

- **Wired Internet Experience (Low vs. High):** The acceptance of mobile technologies as an alternative business and entertainment paradigm in itself depends on how experienced the public is with wired Internet services. In Japan, due to high costs for wired telephone and Internet connections, wireless Internet has become the primary means of communication. But in other parts of the world, such as Western Europe, the wireless Internet has been marketed as only an alternative means of accessing the Web, not as the next natural progression of the cyber world.
- **Billing Scheme (Centralized vs. Decentralized):** In Japan, there exists a central iMode site (the Japanese equivalent of European WAP services) that provides a centralized billing facility, where the users are charged for value-added services via their phone bills. The WAP users, on the other hand, have to pay beforehand with their credit cards for services they want to use, as no such centralized billing system is available in Europe.
- **Pricing (Moderate vs. Medium):** Pricing differences between wired Internet and wireless services also have led to the successful acceptance of mobile phones as the primary means of communication and entertainment in the eastern world. Fixed-line connections are priced at a much higher rate than wireless connections there. The situation is just the opposite in Europe where the mobile industry is yet to be established as the primary mode of communication due to the cheap landline connections.

Figure 2. Factors determining choice of mobile services



STUDYING M-COMMERCE USAGE PATTERNS

The importance of cultural factors affecting and/or determining patterns of m-commerce usage has been emphasised in the earlier sections. Keeping these in mind, the authors conducted a research study in order to investigate the perception and acceptability rates of different mobile services among current/potential users. The study was based primarily in Europe, with about 200 subjects from different communities, age groups, professions, and so forth. However, the teenagers/young adult age group (19-30 years) was chosen as being the major focus group, since this generation is known to have the most familiarity with and access to the latest digital technologies. Some of the research findings are briefly presented below. Although it is hardly a complete discussion of the different dimensions that affect the m-commerce market in Europe, the issues discussed do provide insight into the users' perception of the industry and the services provided.

Factors Determining Choice of Mobile Services

The survey revealed that while the telecommunication giants in the UK are heavily campaigning and gearing up for a 3G mobile services

launch on a global scale, the users still do not see the service package offered as the one major criteria that might influence their choice of operators. While the cost of calls still remains the most important factor behind choosing an operator, other factors such as network coverage are also considered to be more vital than the range of services offered. Figure 2 shows the users' preferences.

Customer Saturation

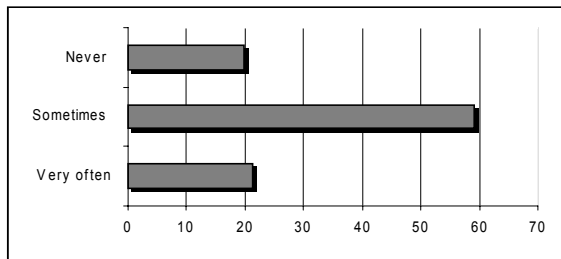
Another factor that has often been described as a major concern of the m-commerce industry not having taken off as promised is the aspect of *customer saturation*. While mobile companies launch new service packages and upgrades on a frequent basis, a very large segment of the customer population just does not feel inclined to possess the latest in the market. The reasons for this are twofold.

Firstly, mobile phone usage has a very high penetration rate in European markets. A substantial percentage of the customer base already possess a mobile phone of fairly high capacity, and they are satisfied with the same. This prevents a constant lookout for latest upgrades.

The second reason, which is more fundamental, is that unlike the Japanese society, a mobile phone, although highly essential, is not perceived as a fashion statement in Europe.

Figure 3. Customer demand for latest services offered/upgrades

(Question: Are you on the constant look out for better service package/upgrade opportunities?)



The possession of the latest in the market is therefore rendered unnecessary.

Only about 21% of the users in the research answered in favour of a constant lookout for latest upgrades. Given that the focus of this research was the *techno savvy* younger generation, this is an alarmingly low rate. The reason for this can be traced back to the how current/potential users *perceive* the mobile industry. Albeit the latest, the services offered are still considered to be *additional/extra*. Unless this changes to *necessary*, it is difficult for the mobile industry to realise its full potential. Figure 3 illustrates the related findings from the research.

Education as a Determining Factor

Another interesting fact that emerged out of the research was the co-relation between a person's

experiences of mobile technologies with his/her education level. More than 90% of the users who use their mobile phones on a regular basis (not just for making calls, but also availing other services offered) are educated to at least "A" levels or equivalent. The importance of this also lies in the fact that it might be a useful strategy for generating personalized contents. An individual's educational background is likely to have a high impact on his/her choices, tastes, and preferences. Thus it might be possible to customize the content in order to reflect the choices of people with different educational backgrounds. It can also be inferred that an individual's educational background would affect his/her social status as well, which in turn is an important determinant of personalized, customized contents. Figure 4 illustrates this co-relation, as revealed in the research.

A Comparative Evaluation of the Different Services Provided

The range of different services that m-commerce offers also presented an interesting theme for the study. These services can be broadly divided into the following categories:

- Information Services
- Entertainment Services
- Financial Transactions
- Location-Based Services

Figure 4. Education levels of m-commerce users

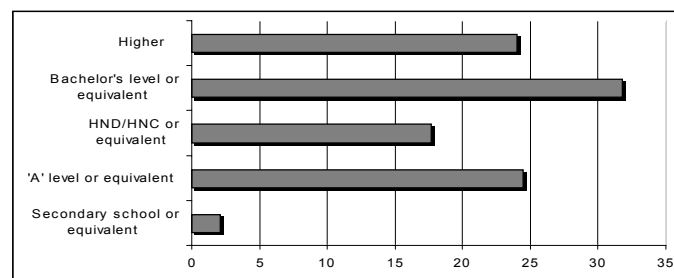


Table 1. Ranking m-commerce services (1—most important; 4—least important)

Rank Value	1		2		3		4	
	Response Frequency	Percentage	Response Frequency	Percentage	Response Frequency	Percentage	Response Frequency	Percentage
Information services	134	69.79	33	17.46	13	6.95	12	6.35
Entertainment service	22	11.46	53	28.04	50	26.74	64	33.86
Financial Transactions	15	7.81	43	22.75	67	35.83	63	33.33
Location based services	21	10.94	60	31.75	57	30.48	50	26.46

While this is an important input to design content strategy formulation, it might also be an important issue to consider for developing marketing strategy.

The research findings show that in spite of all the marketing hype about the range of services that m-commerce promises to provide, almost 70% of people still ranked traditional information services as most important (see Table 1). Both entertainment services and financial transactions have been ranked as the least important by one-third of the respondents.

Cost of Services

In spite of the advanced technologies, user-friendly nature, and exciting applications provided, costs still remain the primary criteria behind the users subscribing to such services. While some services such as ring tone/music download have gained quite some popularity among the public due to relatively cheap rates,

other large applications are priced at a comparatively higher rate. Moreover, in case of some applications or service providers, certain hidden costs are also involved. This includes paying an additional per-use fee, registration charges, and so forth. The cheaper alternative of the wired PC-based Internet thus gains preference over the wireless Internet. So it was not surprising when more than 70% of the subjects interviewed said that they are likely to consider the services offered if the costs involved are less (see Figure 5).

In relation to the above, it was also interesting to note the subjects' estimated monthly budget for their mobile phone services. While on one hand, the young adults have been determined as the most prolific users of m-commerce, it was also found that they have a not-so-high budget, thus preventing them from subscribing to the services available. More than 50% of the people interviewed had a monthly budget of £15 or less, while another 20% had a

Figure 5. Determining the likelihood of users subscribing to cheaper services

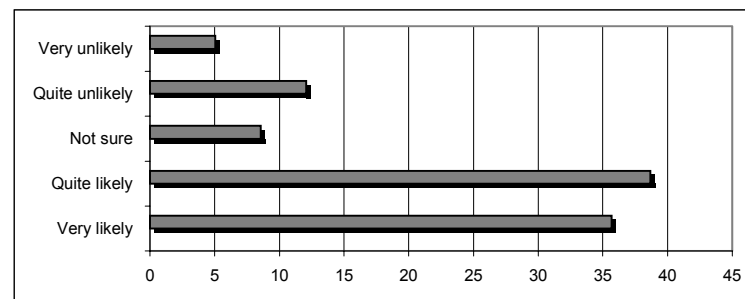
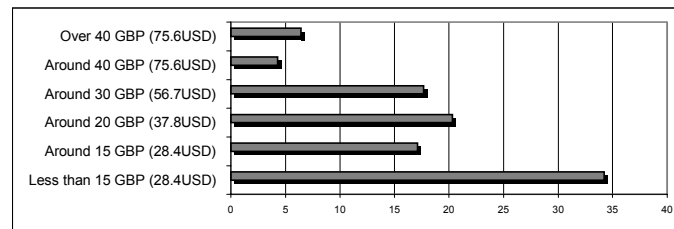


Figure 6. Estimated monthly budget for your mobile phone service (including both voice telephony and data service) (1GBP=1.89USD on March 2005)



budget of £20. Thus marketing different s-commerce services to the young-adults user group, in spite of them being the most technology enthusiastic generation, is placed under certain constraints. The relevant research findings are shown in Figure 6.

CONCLUSION AND FUTURE DIRECTIONS

As the above statistics and discussions show, the market for m-commerce in Europe is yet to fulfil its initial promises. The picture in other parts of the world is not perfect either. The United States, as discussed earlier, has not been too successful in its effort to establish a successful m-commerce infrastructure and industry. The only country to have been able to meet the expectations is perhaps Japan, where telecommunications market works in close contact with their equipment makers, as well as the government—that is, the communications market in Japan, as well as in Hong Kong and Korea, is highly regulated by the government.

It is difficult to identify a unique set of reasons acting as an impediment to the successful realisation of the promises made in the early days of m-commerce. While the technology is definitely in place and ever improving, the focus has rightly shifted to a socio-cultural analysis of the situation over the past few years.

In this chapter, an effort has been made to present a complete overview of the m-commerce industry. While the initial sections have outlined the technological infrastructure and the applications/services offered, the later ones have discussed the market structure and its determinants. However, the most important aspect of this exercise has been the attempt to shift the focus of analysis from a technology-driven perspective to a socio-cultural one. However, even in a socio-cultural mode, it is important to make a conscious effort to refrain from a reductionism way of analysis. It must be remembered that the objective is not to decipher a certain number of explanations, but rather endeavouring to draw the reader's attention to a large plethora of possible social-cultural, political, and economic dimensions that might have contributed towards the not-so-successful m-commerce industry in Europe.

It is also very difficult, if not impossible, to determine the future of the industry. While it is almost certain that the industry did go through a period of over-hype in its early days, there is also no denying of the fact that the opportunities presented to the human civilisation in this m-commerce age were hitherto unseen, and they indeed mark the beginning of a new chapter in the world of digital technologies. As Jarvenpaa et al. (2003) write: "M-commerce lets the Web come to the user, at any time, any location."

However, the m-commerce industry still lacks the vital quality of being *necessary* rather than its present status of just an *additional* set of services. The main challenge that the industry faces, which in turn also determines future research directions in this area, lies not in improving and refining the technology, but rather in changing the users' perception of the services provided as the services being something that will positively affect their lifestyle and surroundings. Such a change in perception, which will take m-commerce growth to a new level, can only be brought by an appreciation of the various socio-cultural dynamics that influence the consumers' choices, beliefs, and attitudes.

REFERENCES

- 3G Newsroom.com. (2003, October 9). *South Korea claims success with 3G*. Retrieved March 30, 2005, from http://www.3gnewsroom.com/3g-news/oct_03/news_3831.shtml
- Balasubramanian, S., Peterson, R. A., & Jarvenpaa, S. L. (2002). Exploring the implications of m-commerce for markets and marketing. *Journal of the Academy of Marketing Science*, 30(4), 348-361.
- Baldi, S., & Thaung, H. P. (2002). The entertaining way to m-commerce: Japan's approach to the mobile Internet—a model for Europe? *Electronics Market*, 12(1), 6-13.
- Barnes, S. J. (2002). The mobile commerce value chain: Analysis and future developments. *International Journal of Information Management*, 22(2), 91-108.
- Borland, J. (2002). *WorldCom piles on telecom collapse*. CNET News.com. Retrieved March 30, 2005, from http://news.zdnet.com/2100-1009_22-939488.html
- Cherry, S. M. (2004, March). WiMax and Wi-Fi: Separate and unequal. *IEEE Spectrum*, 43(3), 16.
- Coursaris, C., & Hassanein, K. (2002). Understanding m-commerce. A consumer-centric model. *Quarterly Journal of Electronic Commerce*, 3(3), 247-271.
- Evans, R. (2004, December 9). *Mobile phone users double since 2000*. Retrieved March 30, 2005, from <http://www.computerworld.com/mobiletopics/mobile/story/0,10801,98142,00.html>
- Ferguson, R. W. Jr. (2002, May 9). Implications of 9/11 for the financial services sector. *Proceedings of the Conference on Bank Structure and Competition*, Chicago, IL (pp. 46-52).
- Granger, V., & Huggins, K. (2000, June). *Wireless Internet—more than voice: The opportunity and the issues*. Report, Merrill Lynch Global Securities Research and Economics Group, USA.
- Gunasaekaran, A., & Ngai, E. (2003). Special issue on mobile commerce: Strategies, technologies and applications. *Decision Support Systems*, 35(1), 187-188.
- Hall, J. (2001, February 17). Chatty teens seen as growth market for wireless. *Reuters Technology News*.
- Hofstede, G. (1980). *Culture's consequences*. London: Sage.
- Hutchison Whampoa. (2004). *Audited results for the year ended 31 December 2003*. Retrieved March 30, 2005, from <http://www.hutchison-whampoa.com>
- Jarvenpaa, S. L., Lang, K. R., Takeda, Y., Tuunainen, K. V. (2003). Mobile commerce at crossroads. *Communications of the ACM*, 46(12), 41-44.

- Kwok, B. (2004, January 3). Watershed year for mobile phones. Companies and finance in South China. *Morning Post* (Hong Kong).
- Lamont, D. (2001). *Conquering the wireless world: The age of m-commerce*. Oxford: Capstone.
- Leung, K., & Antypas, J. (2001). Improving returns on m-commerce investment. *Journal of Business Strategy*, 22(5), 12-14.
- Mahatanankoon, P., Wen, J., & Lim, B. (2004). Consumer-based m-commerce: Exploring consumer perception of mobile applications. *Computer Standards and Interfaces*. Retrieved from www.sciencedirect.com
- Mizukoshi, Y., Okino, K., & Tardy, O. (2001). Lessons from Japan. *Telephony*, 240(3), 92-95.
- NTT DoCoMo. (2004). *Results for the third quarter of the fiscal year ending March 31, 2004*. Retrieved March 30, 2005, from <http://www.nttdocomo.co.jp>
- Shim, J. P., & Shim, J. M. (2003, September/October). M-commerce around the world: Mobile services and applications in Japan, Korea, Hong Kong, Finland, and the U.S. *Decision Line*, 34(5), 9-13.
- Simon, S., Brilliant, P., Macmillan, R., et al. (2000, September). *Wireless Internet report: Boxing clever*. Morgan Stanley Dean Witter.
- Straub, D. (1994). The effects of culture on IT diffusion: E-mail and fax in Japan and the U.S. *Information Systems Research*, 5(1), 23-47.
- Tsalgatidou, A., & Veijalainen, J. (2000, September). Mobile electronic commerce: Emerging issues. *Proceedings of EC-WEB 2000, the 1st International Conference on E-Commerce and Web Technologies*, London, Greenwich, UK (pp. 477-486).
- Wireless Week. (2004). *Buying numbers*, p. 30.
- Wu, J., & Wang, S. (2004). What drives mobile commerce? An empirical evaluation of the revised technology model. *Information and Management*. Retrieved from www.sciencedirect.com

Chapter L

The Mobile Network as a New Medium for Marketing Communications: A Case Study

Heikki Karjaluoto

University of Oulu, Finland

Matti Leppäniemi

University of Oulu, Finland

Jari Salo

University of Oulu, Finland

Jaakko Sinisalo

University of Oulu, Finland

Feng Li

University of Newcastle upon Tyne, UK

ABSTRACT

This chapter discusses the mobile network as a new medium for marketing communications. It illustrates that the mobile medium, defined as two-way communications via mobile handsets, can be utilized in a company's promotion mix by initiating and maintaining relationships. First, by using the mobile medium companies can attract new customers by organizing SMS (short message service) -based competitions and lotteries. Second, the mobile medium can be used as a relationship building tool as companies can send information and discount coupons to existing customers' mobile devices or collect marketing research data. The authors explore these scenarios by presenting and analyzing a mobile marketing case from Finland. The chapter concludes by pondering different future avenues for the mobile medium in promotion mix.

INTRODUCTION

In the repercussion of the mobile hype around wireless access protocol (WAP), followed by the launch of third-generation (3G) networks/ Universal Mobile Telecommunications System (UMTS), the debate over the role of the mobile medium in promoting goods and services has emerged as a topic of considerable magnitude that echoes across different academic disciplines. The burst of the telecommunications bubble in 2000 eventually led telecommunications companies and information technology firms to change their way of thinking, from a technology-driven viewpoint to a more user-oriented perspective. In Europe, only a few mobile services have prospered, while others like many WAP-based services have proved to be unpopular (e.g., Williams, 2003). In fact, only ring tone downloading, logo services, and Short Message Service (SMS) can to date be considered as successful mobile services. The reasons underlying the success of these services fundamentally lie with the strong market demand and easy-to-use technology. When thinking about future mobile services, the *Mobile Internet* is often seen as a messiah of the 3G. Third-generation mobile telephony protocols support higher data rates, measured in kbps (kilobits per second) or Mbps (megabits per second), intended for applications other than voice-centric (3GPP, 2005; Symbian Glossary, 2005). The underlying idea of the 3G/ UMTS networks is that mobile phones are always connected to the best available network ranging from 2G GSM networks to EDGE (General Packet Radio Service), HSCSD (High-Speed Circuit-Switched Data) to WLAN (Wireless Local Area Network), and 3G networks. However, many companies operating in the telecommunications field are facing the same challenge when thinking about the right mobile services to the right mobile users. Recently, a

project led by Nokia and a couple of other Finnish companies announced that television will find its way on mobile phone screens. Consumer acceptance of mobile TV services as well as the underlying technology will be tested and developed with 500 users in Finland (Nokia, 2005).

Since the future of mobile services is still unpredictable, this chapter will not speculate on new mobile services that might take off in the next few years. Instead, we will focus on technologies and applications that are already here and in use, which allow us to examine the utilization of text messaging (SMS) in managing customer relationships in the business-to-consumer markets. In this chapter, we will present a mobile marketing case in a Finnish general store that integrated mobile media in its marketing communications mix as shown in Figure 1.

BACKGROUND TO THE RESEARCH PROJECT

This research is based on a project called PEAR, Personalized Mobile Advertising Services (www.pear.fi), which aims at developing a multi-channel mobile marketing service system for planning, implementing, and analyzing mobile marketing that utilizes value-added features such as personalization, user grouping, presence, profile, and location information. The service system will be tested and developed with end users in real-life settings. The results are expected to contribute to the invention of new customer-oriented service concepts and business models, which can open up potential new business opportunities in global markets. Mobile marketing is in this project defined as marketing communications sent to and received on smart phones, mobile phones, or personal digital assistants (PDAs).

The Campaign Logic

The basic idea of the advertising campaign was to redirect customers to the company's Web page and to get them to register on the company's electronic marketplace. The campaign was advertised in various media (print media, Web pages, and at the store). Advertisements contained instructions of how to participate in the lottery that offered a prize worth 200EUR for registered users. Users were requested to send a text message to a short number and receive a text message back from the company that contains a five-digit short code. The mobile marketing service system generated 100,000 different five-number digits so each participant received an individual code. With the use of this "lucky number"—the five-digit code—customers were able to register with the online shop and thus participate in the lottery.

INTEGRATING THE MOBILE MEDIUM INTO THE MARKETING COMMUNICATIONS MIX

Generally speaking, marketing communications refer to the promotion of both the organization and its offerings (Fill, 2002). The marketing communications mix, also called the promotional mix, comprises a set of tools that can be

used in different combinations and in different degrees of intensity in order to communicate with a target audience. In recent years, the traditional way of thinking about how firms communicate with their customers has changed (e.g., Duncan & Moriarty, 1998; Kim, Han, & Schultz, 2004). With the help of new technologies, companies nowadays have a variety of digital channels allowing assorted ways to both send and receive information. Broadly speaking, we have been witnessing a change from mass communications to more direct and personal communications in which the messages are highly targeted and personalized. This has happened especially in digital communication channels (Kitchen & Schultz, 1999).

Mobile marketing (m-marketing) communications, defined as all forms of marketing, advertising, or sales promotion activities aimed at consumers (MMA, 2003), is one of the most modern digital channels in the promotion mix. Its role in advertising campaigns has not been studied widely, and relatively little is known of its role in the overall communications mix (Karjaluoto, Leppäniemi, & Salo, 2004). M-marketing can be either *push* based, which refers to communications such as SMS alerts sent to wireless devices requiring user permission, or *pull* based, which refers to information a user requests from a provider or advertiser (Barnes & Scornavacca, 2004; Carat Interactive, 2002). The mobile medium has to date mainly been used in promotions such as lotteries and various competitions (e.g., Pura, 2002). However, the market seems to be ready for more sophisticated two-way mobile marketing campaigns such as mobile customer relationship management (Finnish Direct Marketing Association, 2004), as customers are more and more using mobile data services such as text messaging and multimedia messaging in buying purposes and in providing feedback.

Figure 1. Mobile marketing campaign integrated with other marketing channels

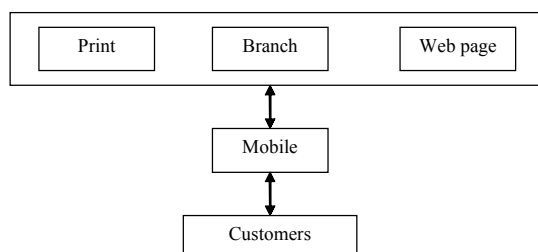
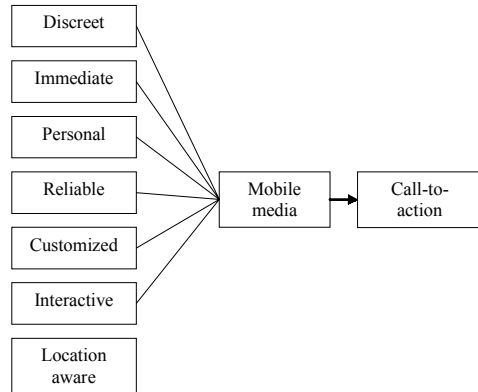


Figure 2. Special features of mobile media



SPECIAL FEATURES OF THE MOBILE MARKETING MEDIUM

The mobile medium has some unique features that other direct marketing channels lack. In general terms, the mobile medium is favoured by marketers for its broad reach, low cost, and high retention rates (Clickatell, 2002). For mobile phones, several features are particularly relevant: the mobile phone is seen as an extremely personal, immediate, and interactive medium allowing marketers an effective way to reach customers in a fresh manner (Koranteng, 2001; Peters, 2002). As a marketing communication channel, the mobile (especially text) messaging is seen as immediate, automated, reliable, personal, discreet, and customized, allowing an efficient way to reach markets directly and providing mobile phone users a direct call-to-action, which would be almost impossible via other channels (Barnes & Scornavacca, 2004; Clickatell, 2004; Leppäniemi & Karjaluo, 2005). These special features are illustrated in Figure 2.

As mobile phones are extremely personal in nature, advertising to mobile devices has to be very discreet in the sense that unwanted messages are easily perceived as spam. Messaging

services (e.g., SMS and MMS) can be considered a very reliable way to distribute information, not only due to the fact that messages almost always arrive in time, but also because the majority of consumers usually read all messages they receive. In relation to mobile marketing campaigns, studies reported that in over 90% of cases, respondents read mobile advertising messages they receive (Enpocket, 2003). Moreover, the mobile media can also be regarded as an interactive media. As Enpocket's (2003) study indicated, of 5,000 consumers participating in SMS campaigns, 15% of them responded to mobile marketing messages. Finally, mobile devices have the ability to identify the location of users through the use of various technologies such as network-based positioning (or remote positioning), accurate local area positioning techniques, and satellite positioning (Kumar & Stokkeland, 2003; Zeimpekis, Giaglis, & Lekakos, 2003). Location-based advertising (LBA) is based on the idea that in a certain location, and additionally at a certain time, consumers receive advertisements based on their location (Salo & Tähtinen, 2005; Tsang, Ho, & Liang, 2004). However, as the rules of protecting consumers' privacy, including the use of location-based information for marketing purposes, are becoming stricter, the development and diffusion of location-based advertising have many obstacles to overcome.

Research has shown that the primary role of mobile marketing in a company's promotion mix has to date been promoting call-to-action (e.g., Paananen, 2003; Clickatell, 2004; Karjaluo et al., 2004). In providing a direct call-to-action, location awareness and time open up the possibility to personalize messages in the manner that provide straight call-to-action. For instance, if a consumer arrives at a store, he or she might receive a personalized advertisement from that store based on his or her profile. It is important to note that to receive the benefits,

for instance a discount to the store, consumers need to be in a certain location.

CUSTOMER RELATIONSHIP MANAGEMENT WITH MOBILE PHONES

Over the past decade it has become increasingly difficult to differentiate from competitors in serving the general product needs of customers. Therefore, companies have had to shift their focus to customer orientation and to search for novel ways to create value to customers. As a result, customer relationship management (CRM) is currently gaining widespread popularity in several disciplines and industries (e.g., Ryals, 2003; Zablah, Bellenger, & Johnston, 2004). On the one hand, the objective of CRM is to build and maintain customer relationships, and on the other to provide value for customers. Despite the potential of traditional CRM to provide value for customers, customers are expecting more and more individual attention. From the viewpoint of marketing communications, new digital marketing channels such as the Internet and mobile phones are considered to be powerful new media to reach consumers by allowing personalization and interactivity of both the content and the context of the message (Heinonen & Strandvik, 2003; Kim et al., 2004). Furthermore, as mobile marketing can combine the capacities of both direct marketing and ever-present nature and power of mobile digital technology, this form of communication is seen to provide synergy that will increase the potential of direct marketing (Mort & Drennan, 2002). Although electronic customer relationship management (e-CRM), defined broadly as CRM through the Internet, has received much attention among practitioners and academics (Bradshaw & Brash, 2001; Feinberg, Kadam, Hokama, & Kim, 2002; Fjermestad & Romano,

2003), the mobile medium as an element of CRM has gained far less attention.

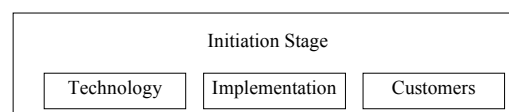
Customer relationship management with mobile phones—in other words mobile customer relationship management (we use the term m-CRM)—can be defined as an ongoing process that provides seamless integration of every area of business that touches the customer, for the purpose of building and maintaining a profit-maximizing portfolio of customer relationships, by taking advantage of the mobile medium.

Because customer relationships evolve with distinct stages (Dwyer, Schurr, & Oh, 1987), companies should also interact with customers and manage relationships differently at each stage (Srivastava, Shervani, & Fahey, 1998). The CRM process outlines three key stages, namely the initiation, maintenance, and termination phases (Reinartz, Krafft, & Hoyer, 2004). In this study, it is implicitly assumed that m-CRM consists of these three stages as well. Because the main interest of this study is about how to redirect customers to the company's Web page and to get them register to the company's electronic marketplace, we focus primarily on the initiation stage.

USING MOBILE CRM IN INITIATING RELATIONSHIPS

Integrating the mobile medium as an element of CRM involves three key aspects: technology, implementation, and customers. This has been illustrated in Figure 3.

Figure 3. Dimensions of the initiation stage of m-CRM



The Mobile Network as a New Medium for Marketing Communications

The first key aspect to consider during the initiation stage of m-CRM is the technology. The company has to decide whether to rent the hosting mobile marketing server or to acquire it. Then gateways have to be built to connect with the mobile operators whose customers are allowed to be contacted.

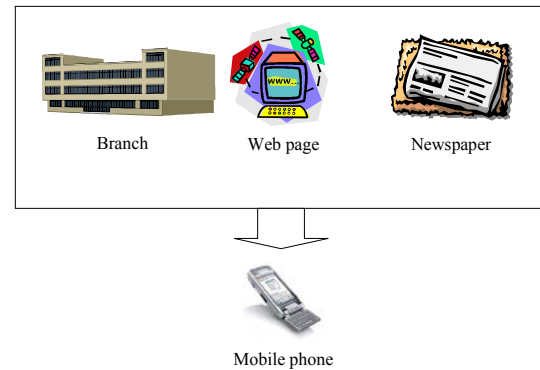
The implementation consists of two decisions: first, through what marketing medium that the customers are acquired; and second, how much it will cost customers to use the messaging service—that is, sending the SMS message.

Given the personal nature of mobile phones, customers are often unwilling to use the mobile medium for marketing purposes due to fears of unsolicited marketing messages or even spam. Therefore, the third aspect of the initiation stage comprises understanding customers; to discover what are the key factors that lure the customer to use the mobile medium; and in this case, to redirect them to the company's electronic marketplace.

CASE STUDY

The central idea of the case study was to analyze the role of mobile media in generating traffic to company A's e-commerce Web page. Moreover, we wanted to investigate the company's (an ironmonger store) consumer responses to mobile marketing. The marketing campaign began at the end of November 2004 with a full-page advertisement in local newspaper. The newspaper advertisement contained instructions on how to participate in the SMS lottery. Consumers were asked to send a simple text message to a five-digit short number code. Immediately after sending the message, the mobile service system generated a four-digit code that the consumer received on his or her phone. This code was then entered on the

Figure 4. Marketing communications mix in company A's case



company's Web site. After entering the code to the Web site, consumers were guided to the registration page and asked to fill in the registration form. In case a registered person is participating in the campaign, they are guided to update their registration information. All registered consumers would be entered into a drawing for a digital video camera. The logic of the campaign is illustrated in Figure 4.

MAIN RESULTS

Altogether 232 consumers took part in the lottery during the period from November 24 to December 20, 2004. The newspaper advertisement on November 24, 2004, generated approximately 100 new registered customers in two days. After the newspaper advertisement, the company's Web site was, practically speaking, the only advertising channel for the lottery. The amount of new registrations per day varied from a few to around 10, with a final number of 232 registered consumers. Of these, only five were existing customers of the case company's online store. There were 47 female (20.3%) and 185 men (79.7%) in the sample. Approximately 33% of the participants were between

ages 36 and 49, 22% were between 26 and 35, and 12% were between 50 and 64. The remaining categories had less than 11% of responses per category. With respect to careers, most respondents were blue-collar workers (34%), followed by white-collar workers (24%); 13% belonged to top or middle management and 11% to lower management, and students were 15%. Additionally, the registration form inquired about respondents' areas of interests. A total of 118 respondents reported interest in electronics and photography, followed by computers, leisure time, and motor vehicles. Of the total 21 interest areas, pets (26 responses), toys (29 responses), and watches and jewelry (30 responses) were considered least interesting.

CONSUMER WILLINGNESS TO RECEIVE DIRECT MARKETING

The registration form finally inquired about respondents' willingness to receive digital marketing communications from the company either via SMS or e-mail. While most respondents wanted to opt out of text messaging service (73%), the majority opted in to e-mail marketing (67%). This finding might relate to the fact that the case in question was relating to online commerce, and thus respondents regarded the e-mail channel as more suitable for them to receive marketing communication. In addition, approximately 24% of the respondents opted out of both marketing communication channels.

We next compared willingness to receive SMS and e-mail marketing with demographic variables with the use of a series of chi-square analyses. In respect of gender, 25% of males welcomed SMS marketing from the company, whereas the corresponding number for females was 34%. Concerning willingness to receive e-mail marketing, around 70% of males and 55%

of females opted in to e-mail marketing. However, the difference between genders was neither statistically significant for SMS (chi-square=1.612, $p=0.204$) nor for e-mail (chi-square=3.510, $p=0.061$), although in e-mail it was relatively close to significant ($p<0.05$ level). With regard to respondents' age and willingness to receive digital marketing communications, a statistically significant difference was found in welcoming e-mail marketing (chi-square=15.725, $p=0.008$). This finding indicates that the younger age categories were not so willing to receive e-mail marketing compared to the older ones. Age category 26 to 35 most eagerly welcomed e-mail marketing (86% opted in), followed by age group 36 to 49 (70% opted in), and age group 50 to 64 (64% opted in). Approximately 50% of those under 18 opted in to e-mail marketing, as the corresponding number for age group 19 to 25 was 48%. No relationship was found between age and willingness to receive SMS communications (chi-square=4.752, $p=0.451$).

In terms of careers and willingness to receive marketing communications, no differences were found with the original classification of professions. However, after recoding the profession variable into a two-class variable, in which 1 was equal to blue-collar worker and 2 equal to white-collar worker, a statistically significant difference was found between profession and willingness to receive SMS communications (chi-square=5.520, $p=0.019$). It seems that blue-collar workers more eagerly welcomed SMS communications (33% opted in) than white-collar workers (15% opted in). No differences were found between profession and willingness to receive e-mail marketing (chi-square=0.236, $p=0.627$).

Finally, there is some evidence that respondents' place of residence has an impact on willingness to receive both SMS and e-mail marketing. The place of residence variable was

divided into three categories, in which one was equal to “I live within ten-mile radius of the company’s physical store,” two was equal to “I live within a hundred-mile radius from the company’s physical branch,” and three was equal to “I live elsewhere in Finland.” Interestingly, with respect to willingness to receive SMS marketing, most respondents living elsewhere in Finland did not want to receive SMS marketing (80% opted out). A total of 71% living near the store opted out of SMS marketing, and 65% of those living in a hundred-mile radius opted out. The difference between groups was close to statistical significance ($\chi^2=5.664, p=0.059$). Moreover, those living the farthest away from the store were most willing to welcome e-mail marketing (75% opted in), followed by those living within a hundred-mile radius (60% opted in), and those living near the store (54% opted in). The χ^2 value ($\chi^2=6.860, p=0.032$) indicates a statistically significant relationship between place of residence and willingness to receive e-mail marketing. On this basis, it seems that the further a person lives, the more likely he or she is to opt in to e-mail marketing. However, in regard to SMS marketing, there is no linear relationship between location and willingness to receive SMS marketing.

To conclude, our results indicate that blue-collar workers are more willing to receive SMS marketing from the case company than white-collar workers. This fresh finding should be interpreted in light of the empirical case setting, which might indicate that blue-collar workers are not in general so familiar with e-mailing as white-collar workers, and thus they regard the SMS channel as more suitable for them in receiving marketing communications. However, validation of further work is needed. Moreover, with regard to welcoming e-mail marketing, it seems that age groups 26-35, 36-49, and 50-64 offer the most potential for e-mail marketing

campaigns, as more than 50% of them welcomed e-mail marketing. Furthermore, it seems that age itself has no influence on willingness to receive SMS marketing communications. Finally, our results showed that place of residence has an impact on willingness to receive e-mail marketing. The further a person lives from the physical store, the more likely he or she was to opt in to e-mail marketing communication.

CONCLUSION AND FUTURE DIRECTIONS

This chapter responds to the call for research on the use of mobile media in the marketing communications mix by investigating its specific features and role in integrated marketing communication mix and by describing the use of mobile media in amassing mobile marketing customer database. Our objective was to discuss the ways to integrate the mobile medium into the promotion mix of companies. By using a single case study from the retail sector, we showed that by combining the mobile medium and e-commerce store, it is possible to build a customer database in an efficient and cost-effective manner. Although our empirical case mainly contributes to the discussion of how to initiate customer relationships, it also gives some insights into the maintenance process of relationships by asking respondents about their digital channel preferences.

In light of the main results, several conclusions can be drawn. First of all, with relatively small promotional activity, the case company gained close to 250 new registered customers to their online store mainly by the use of a newspaper advertisement and online advertising on the company’s own Web site. As most respondents opted in to e-mail marketing (67%), and most opted out of SMS marketing (73%), in

the next stage of relationship building, the e-mail channel might be the right avenue to continue. However, some specific customer segments (blue-collar workers especially) welcomed SMS marketing. This finding should be interpreted as: those customers who opted in to SMS channel either do not use e-mail or for other reasons want marketing communications via SMS. By giving customers a choice, it is supposedly also contributing to the overall customer satisfaction and thus driving those registered potential customers into purchasing online customers as well.

This study has some limitations that present opportunities for further research. First, the study is among the first ones examining the use of a mobile device as a marketing communication channel, and thus the results obtained should be considered tentative. Second, we study only one retailer and its marketing communication, and despite the fact that its communication mix is in line with other companies operating in the same field nationwide, it would be valuable to scrutinize other retailers as well. In sum, we assume that these limitations do not endanger the reliability and validity of the findings, yet they do place bounds on the conclusions and implications that can be drawn from the study.

While mobile marketing is today almost entirely SMS based, the diffusion of MMS-enabled phones will presumably shape the industry in the future (Paananen, 2003; Barwise & Strong, 2002). Also the mushroom of devices with larger screens will guide mobile marketers to new avenues (Yonos, Gao, & Shim, 2003). These should be taken into consideration when planning future studies. Furthermore, a natural extension of our study would be the investigation of the role of mobile media in the marketing communications mix with other retailers. By doing so we might get valuable insights into how companies nationwide use, or plan to use, mobiles as a media in marketing communications.

REFERENCES

- 3GPP. (2005). *Terms & abbreviations*. Retrieved March 17, 2005, from <http://www.3gpp.org>
- Barnes, S., & Scornavacca, E., (2004). Mobile marketing: the role of permission and acceptance. *International Journal of Mobile Communications*, 2(2), 128-139.
- Barwise, P., & Strong, C. (2002). Permission-based mobile advertising. *Journal of Interactive Marketing*, 16(1), 14-24.
- Bradshaw, D., & Brash, C. (2001). Managing customer relationships in the e-business world: How to personalize computer relationships for increased profitability. *International Journal of Retail & Distribution Management*, 29(12), 520-529.
- Carat Interactive. (2002). The future of wireless marketing. *Research Report*. Retrieved from http://www.mmaglobal.com/resources/Wireless_WhitePaper.pdf
- Clickatell. (2004). Business mobility guide. *Research Report*. Retrieved from <http://www.clickatell.com>
- Duncan, T., & Moriarty, S. E. (1998). A communication-based marketing model for managing relationships. *Journal of Marketing*, 62(2), 1-13.
- Dwyer, R., Schurr, P., & Oh, S. (1987). Developing buyer-seller relations. *Journal of Marketing*, 51(2), 11-28.
- Enpocket. (2003, February). The response performance of SMS advertising. *Research Report*, 3. Retrieved from <http://www.enpocket.com>
- Feinberg, R. A., Kadam, R., Hokama, L., & Kim, I. (2002). The state of electronic cus-

- tomers' relationship management in retailing. *International Journal of Retail & Distribution Management*, 30(10), 470-481.
- Fill, C. (2002). *Marketing communications—Context, strategies and applications*. Englewood Cliffs, NJ: Prentice-Hall, Harlow.
- Fjermestad, J., & Romano, N.C. (2003). Electronic customer relationship management. Revisiting the general principles of usability and resistance—An integrative implementation framework. *Business Process Management Journal*, 9(5), 572-591.
- Heinonen, K., & Strandvik, T. (2003, May 22-23). Consumer responsiveness to mobile marketing. Paper presented at the *Proceedings of the Stockholm Mobility Roundtable*, Stockholm, Sweden.
- Karjaluoto, H., Leppäniemi, M., & Salo, J. (2004). The role of mobile marketing in companies' promotion mix. Empirical evidence from Finland. *Journal of International Business and Economics*, 2(1), 111-116.
- Kim, I., Han, D., & Schultz, D.E. (2004). Understanding the diffusion of integrated marketing communications. *Journal of Advertising Research*, 44(1), 31-45.
- Kitchen, P., & Schultz, D. (1999). A multi-country comparison of the drive for IMC. *Journal of Advertising Research*, 39(1), 21-38.
- Koranteng, J. (2001). ZAP! There's no escaping the mobile ad. *Ad Age Global*, 1(5), 9.
- Kumar, S., & Stokkeland, J. (2003). Evolution of GPS technology and its subsequent use in commercial markets. *International Journal of Mobile Communications*, 1(½), 180-193.
- Leppäniemi, M., & Karjaluoto, H. (2005). Factors influencing consumer willingness to accept mobile advertising. A conceptual model. *International Journal of Mobile Communications*, 3(3), 197-213.
- MMA (Mobile Marketing Association). (2003). *MMA code for responsible mobile marketing*. Retrieved December 14, 2003, from www.mmaglobal.co.uk/
- Mort, G., & Drennan, J. (2002). Mobile digital technology: Emerging issues for marketing. *Journal of Database Marketing and Customer Strategy Management*, 10(1), 9-23.
- Nokia. (2005). *Mobile TV pilot begins in Finland*. Retrieved March 17, 2005, from http://press.nokia.com/PR/200503/983665_5.html
- Paananen, V. M. (2003, February 12). *European mobile marketing: Case Add2Phone*. Presentation at the NETS Seminar, Helsinki, Finland.
- Peters, B. (2002, November 6). The future of wireless marketing. *Carat Interactive Study*. Retrieved from <http://www.caratinteractive.com/resources/articles.html>
- Pura, M. (2002). Case study: The role of mobile advertising in building a brand. In B. E. Mennecke & T. J. Strader (Eds.), *Mobile commerce: Technology, theory and applications* (pp. 291-308). Hershey, PA: Idea Group Publishing.
- Reinartz, W., Krafft, M., & Hoyer, W. (2004). The customer relationship management process: Its measurement and impact on performance. *Journal of Marketing Research*, 41(3), 293-305.
- Ryals, L. (2003). Making customers pay: Measuring and managing customer risk and return. *Journal of Strategic Marketing*, 11(3), 165-175.
- Salo, J., & Tähtinen, J. (2005). Retailer use of permission-based mobile advertising. In I.

- Clarke, III, & T. B. Flaherty (Eds.), *Advances in electronic marketing* (pp. 139-155). Hershey, PA: Idea Group Publishing Group.
- Srivastava, R., Shervani, T., & Fahey, L. (1998). Marketing based assets and shareholder value: A framework for analysis. *Journal of Marketing*, 62(1), 2-18.
- Symbian Glossary. (2005). Retrieved March 17, 2005, from <http://www.symbian.com/technology/glossary.html>
- Tsang, M. M., Ho, S.-C., & Liang, T.-P. (2004). Consumer attitudes toward mobile advertising: An empirical study. *International Journal of Electronic Commerce*, 8(3), 65-78.
- Williams, C. (2003, March 17). Can GPRS get back on track? *Lucent Technologies: 3G Solutions for Operators*, 12. Retrieved from http://www.lucent.com/livelink/0900940380033c60_Newsletter.pdf
- Yonos, H. M., Gao, J. Z., & Shim, S. (2003). Wireless advertising's challenges and opportunities. *IEEE Computer*, 36(5), 30-37.
- Zablah, A. R., Bellenger, D. N., & Johnston, W. J. (2004). An evaluation of divergent perspectives on customer relationship management: Towards a common understanding of an emerging phenomenon. *Industrial Marketing Management*, 33(6), 475-489.
- Zeimpekis, V., Giaglis, G. M., & Lekakos, G. (2003). Towards a taxonomy of indoor and outdoor positioning techniques for mobile location-based applications. *ACM SIGecom Exchanges*, 3(4), 19-27.

Chapter LI

Overview and Understanding of Mobile Business in the Age of Communication

Joseph Barjis
Georgia Southern University, USA

ABSTRACT

This chapter provides an introduction, review and study of mobile businesses with emphasis on its supporting mobile technologies and wireless networking. The chapter first discusses the concept of mobile business where opportunities, motivations and needs for this type of business are studied. Following this discussion, the chapter studies the current status of mobiles business, key hardware and software solutions (business applications) available on the market. The chapter also discusses different mobile devices, communication infrastructure, supporting networks and other crucial components that make businesses mobile and able to be conducted anytime and anywhere. Finally, an extended discussion is focused on issues and future developments of mobile businesses along with some recommendations, and suggestions regarding mobile business.

INTRODUCTION

Mobile business is considered an offspring of the Advanced Communication Age and a driving force of the new economy, therefore its discussion is strongly aligned with the discussion of underlying information and communication technologies (ICTs).

The amazing pace of innovations in ICTs during recent years has opened a wide spec-

trum of new opportunities and challenges for the business industry. These opportunities demand a dramatic shift towards mobility in almost every aspect of life such as education, entertainment, health care, and business. Rapid developments in wireless communication technologies, mobile devices, high-speed transmission facilities, and broad bandwidth prepared and paved the way for transforming human activities towards mobility. The most notice-

able impact of these evolving technologies can be seen in business, which is preparing for another revolutionary change. First, business has gone through transformation from traditional business to electronic business, and now it has been adapting towards mobile business, or m-business for short.

In the coming years, m-business will be a fashion for industry, researchers, enterprise managers, and society as a whole. This will be the business style of the Age of Communication, inspiring managers and enterprises for serious shift. Perhaps even, enterprises will experience another era of Business Process Reengineering, or maybe this time, m-Business Process Engineering (mBPR)!

Before diving deep into the topic, it is worthwhile to mention some example where mobility has been making its initial breakthrough. These different examples aim to provide an idea about the breadth, depth, and diversity of mobility and mobile business.

Education

Educational institutions have been pioneering implementation of the wireless networking environment, providing students with flexibility of accessing campus resources and downloading academic applications at their convenience and desired location (lab, classroom, library, cafeteria, campus garden, or while watching campus games). Not being tied to lab hours and classrooms, students are given more flexibility and opportunity to pursue their education, which in turn increases quality of education. So, campuses are going mobile within the campus area.

Health Care

Hospitals in general and modern medical practices in particular are adapting towards mobile health care delivery. Computer-based patient

records, also referred to as electronic medical records, are part of a system that provides a mobile working environment for physicians, staff, and managers of medical practices. Each physician carries a handheld computer that access patients records, X-rays, and surgery videos, allowing sharing and discussing of images with specialists from other hospitals, coordinating remote operations, and so forth. If a physician wants immediate information about a particular medicine prior to issuing a prescription, his handheld computer allows him to access the relevant Internet page for such information. Productivity and quality of health care service is impressive with mobile facilities.

Sales and Marketing

Retail, wholesale, mass distribution centers are using a mobile business environment for goods delivery, shelf refill, inventory control, warehouse management, transport and logistics, and working with branches in different locations. Mobile devices help to track goods delivery and movement of products.

These three small examples illustrate different ranges of mobility within a building, within an enterprise, and within a town, used for a wide range of activities. These examples help analysts extract some important characteristics of m-business including range of functionality and types of mobile devices (wireless laptops, tablet PCs, smart phones, etc.). According to some authors, the application of m-business can be distinguished as “macro” applications in outdoor settings or “micro applications in indoor environments—for example, hospital, libraries, hypermarkets. Like the underlying wireless networks supporting it, m-business may be distinguished by its span as a local, regional, or global m-business.

With this brief introduction, this section is concluded and the rest of the chapter will

discuss different aspects of m-business and its technological components, and elaborate on various facets of m-business.

CONCEPT OF M-BUSINESS

In general, the main concept of m-business is moving enterprises' critical business to the point of sale and service, or even closer to the point of consumers.

Like its predecessor, electronic business, the concept of mobile business has been used in many applications from communication to consumer transactions and corporate services (Vos & de Klein, 2002). However, the real potential of m-business is much broader than merely providing service, conducting sales, or delivering goods. A well-engineered, well-designed, and well-integrated m-business supports not only conducting business, but also adds collaboration, coordination, instant communication, and management features in the business. Being based on most advanced ICTs, m-business aims to be more productive than traditional business or business supported by networked computers, and productivity means in all aspects.

Evolving from mainframe and wired network eras, m-business is the leading edge of the new generation (3G or third generation) of wireless networking that aims to adapt the best business and management practices, standards, and styles.

Motivations for M-Business

On the business horizon, the main motivations that have been pushing business towards m-business are competition for flexibility in conducting business, extending functionality and service to the business point, convenience of employees and comfort of consumers, better

satisfaction, quality improvements, personalization, and localization of business. And obviously, revenue increase and market gain are at the very core of these driving forces.

On the technology horizon, the revolutionary progress in wireless networking, information technologies, and mobile devices making high-speed wireless communication everywhere has provided new infrastructure for business that resulted in more mobility.

Definition of M-Business

Mobile business is the business of the future which is based on a wireless infrastructure, using mobile devices, bringing critical business to the point of service and sale, aiming for more productivity in wide economic sense. Depending on a focal point and perspective, m-business can be defined in quite different ways.

Mobile business study and application attracted the attention of many outstanding authors, scholars, and researchers (e.g., Deitel, Deitel, & Steinbuhler, 2001; Deitel, Deitel, Nieto, & Steinbuhler, 2003; Vos & Klein, 2002; Paavilainen, 2002). In the abundance of available definitions used by different authors, it is quite challenging to find a unique definition, however the definition given by Kalakota and Robinson (2001) could be cited here as an example: m-business is "the application infrastructure required to maintain business relationships and sell information, services, and commodities by means of the mobile devices." This is just one of definitions that characterizes m-business. Different definitions are given from different perspectives. However, regardless of definition perspective, what is generic about m-business is that m-business encompasses three essential components: wireless networking technology (3G Networks, WLAN, WWAN), mobile devices, and improved business practices (procedures). The last component is a key

component of m-business. If the first two are facilitators, the third one is the main objective of m-business.

M-BUSINESS TODAY

Although application of m-business started not too long ago, this opportunity very quickly attracted enterprise managers, industry leaders, and researchers and authors. Today a number of periodicals are adapted or founded on m-business, numerous monographs are published (e.g., Paavilainen, 2002; Vos & Klein, 2002; Kalakota & Kurchina, 2004; Sadeh, 2002), annual conferences such as ICMB (International Conference Mobile Business) are held, and a wealth of Internet-based resources reporting studies, results, examples, and models of m-business are available. In addition, tens of IT and business consulting companies refocused their activity from electronic commerce and electronic business towards m-business.

Studying the overwhelming opportunities and increasing demands in m-business, Kalakota (2005), in his work “Mobile Business: Vision to Value,” emphasizes how rapidly the emerging technologies change the way enterprises conduct their business and how dramatically m-business is replacing traditional business. As a result of these changes, the author states, questions have shifted from “Should I do mobile business?” to “How can mobile create business value?”

Enterprises and businesses worldwide are implementing mobile business solutions to accelerate business cycles, increase productivity, reduce operating costs, and extend their enterprise infrastructure.

The need to go mobile turned into serious competition between leading companies that provide wireless infrastructure, application solutions, and mobile devices for m-business.

Today, the following market leaders are among the top providers of applications and wireless networking infrastructure for mobile businesses.

- In the *software market*, SAP as the world’s largest inter-enterprise software company took a pioneering initiative in providing software packages for different types of m-business as listed below. In gaining a leading position in mobile business solutions, Microsoft is adding m-business features to Windows.
- In the *wireless market*, Cingular and Verizon are pioneering in providing modern wireless networking services by introducing and expanding 3G networks in major cities and metropolitans (Segan, 2005).
- In the *mobile devices market*, Siemens, Nokia, and other leading providers are introducing mobile technologies that significantly boost competitiveness of businesses.

The SAP mobile business solution set includes ready-made applications that provide access to the corporate information and processes—anytime, anywhere—allowing use of a variety of mobile devices. Among various software packages, SAP provides mobile business applications (SAP, 2005):

- **SAP Mobile Time and Travel:** This package gives mobile workers access to timesheets and travel management functionality.
- **SAP Mobile Sales:** This package provides a solution for salespeople who need to perform their tasks quickly and productively.
- **SAP Mobile Service:** This package enables field service engineers to react quickly to customer needs.

- **SAP Mobile Asset Management:** This package allows in-house service engineers to access relevant business processes anywhere, anytime.
- **SAP Mobile Procurement:** This package enables mobile workers to manage the entire procurement process, from price comparison to ordering.

INFRASTRUCTURE OF M-BUSINESS

Think of m-business as a two-level framework (as shown in Figure 1), where the upper level is the business level and the bottom level is the IT infrastructure that supports business to carry out its mission and tasks, and sets business in motion.

Just in a narrow sense, let us say, from a high-level perspective, one can consider that the business level is a variable and the IT level is relatively a constant. Business level is variable because it represents any type of business, but the IT infrastructure, in general functionality, may be the same for almost any type of m-business, with some differences such as range, size, structure, configuration, potential, and complexity.

The IT infrastructure of m-business includes two major components—mobile devices and wireless networks.

Mobile Devices

One of the main characteristics of mobile devices used in e-business is the communication or networking facility and capacity to receive,

transmit, and process different types of data (text, audio, video) at a high rate. In addition to high-speed data exchange, ease of handling, portability, and size are important in these devices. Furthermore, these devices must be able to download essential business, office, and enterprise applications and have sufficient memory to run them.

Currently, devices used in m-business are wireless handheld computers, laptops, PDAs, tablet PCs, smart phones, Blackberry, and so on. These devices should be capable to run high-performance business, office, and enterprise applications such as multimedia, full-motion video; wireless teleconferencing; and use connection over wireless networks using Wi-Fi, GPRS, Bluetooth, or other advanced connections.

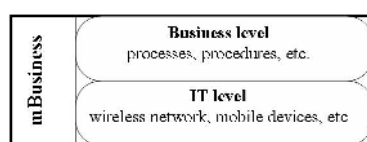
For a better idea and distinction between ordinary mobile devices and business-quality devices, and a better idea about business quality mobile devices, have a look at the features provided by the Nokia 9500, having a full set of critical business tools, full keyboard, with a large memory capacity and versatile network connections:

Browse the Internet in rich full color, on a wide, easy-to-read screen. Work with office documents—not just e-mail and memos, but presentations and databases too. Get them from your corporate network via Wireless LAN or EDGE for fast mobile access. Keep your Personal Information Management data in sync and up-to-date calendar and contacts—with PC Suite and SyncML, you can easily exchange data between your Nokia 9500 Communicator and a compatible PC. (Nokia, 2005)

Wireless Networks

Although not yet in mass application, in the future m-business will be operating on third-

Figure 1. M-business IT infrastructure



generation (3G) wireless networks that provide high-speed download and upload rate. The speed of transmission in these networks using mobile devices is at the level of DSL connections. Because data costs on 3G networks are lower than on traditional networks (Solheim, 2005), more and more enterprises will shift into using 3G networks as a main infrastructure.

Currently, m-businesses are based on different types of networking technology:

- **Personal Area Network:** Using Bluetooth technology enables short-range device-to-device wireless connections within a small space (office, a desktop, a personal space).
- **Wireless Local Area Networks:** Using Wi-Fi technology, based on IEEE 802.20, 802.11 standards, support a wireless connection to a network from inside a home or from a hotspot in a building, campus, or airport.
- **Wireless Metropolitan Area Networks:** Using WiMax, based on IEEE 802.16 standard, will enable any remote worker to make a wireless connection anywhere in a range up to 50 kilometers.
- **Wireless Wide Area Networks or 3G Networks:** Provide the highest available bandwidth for mobile devices. Although theoretical rate of transmission is 2Mb, for practical purposes the transmission speed is like DSL letting users download text, audio, video, Web contents, and e-mail while in motion.
- **4G Technologies:** Promise to integrate different modes of wireless communications, from indoor networks such as wireless LANs and Bluetooth, to cellular signals, to radio and TV broadcasting, to satellite communications.

ISSUES OF M-BUSINESS

In the previous sections we discussed advantages and benefits of m-business; however, enterprise managers and business owners ought to be aware of issues and challenges before undertaking the transformation initiative.

It would be not less than an illusion, if one considered the transition from traditional or e-business to m-business as just shifting from a wired environment to wireless communication and networking. Transformation into mobile business is not about moving from wired environment to wireless, from desktops to handheld computers, or from office to field. And the challenge is not only about acquisition and implementation of best wireless technology or awareness of emerging mobile technology; it is rather a multi-dimensional issue where technology is only one facet of it. Challenges that may require more profound study are transformation of the business and enterprise, shift in the mindset from thinking of m-business as a different way of doing business to considering m-business as an implementation of best business practices, improved procedures, higher quality, and so forth. The challenge of going mobile is much more complicated than application and implementation of mobile technology.

Although each category of business may have its own peculiarities, some of the common issues for enterprises to carefully study while embarking on competing for faster implementation of m-business are as follows

- From business's perspective
 - How well are m-business opportunities studied?
 - How much Business Process Reengineering, ERP system changes, and customer relationship management is required to go mobile?

- What is expected from m-business, direct profit, or quality of service? How can qualitative values be turned into quantitative values?
- How much patience is needed before harvesting the first fruits of benefit?
- Are the employees ready to go mobile, or will they resist against?
- Also important to remember: Mobile business should not be considered as an immediate way for profits.
- From technology perspective
 - What will happen with the existing IT infrastructure?
 - How carefully are the problems of interfacing, integration, and legacy systems studied?
 - While going mobile, an important challenge is security of m-business—is this issue studied?
 - How well are the connectivity and management of mobile devices, security, and updates issues studied?
- From consumer's perspective:
 - What is the impact of m-business on consumers?
 - Is the transition for consumers straightforward or painful?
- Transition cost
 - Should the business mobilize a few employees or the whole enterprise?
 - How much it will cost for an enterprise with tens of thousands of employees?
 - What is the price and benefit of a mobile employee (m-employee)?

For each of the mentioned categories, the list of such questions can be much longer than shown here. These are just some of the issues not including public, political, and legislative issues.

CONCLUSION

This chapter provided a brief overview of m-business, opportunities that m-business opens, and challenges accompanying these opportunities.

The chapter also discussed m-business in connection with its underlying technology. In this part different technological components of m-business were introduced and discussed. Along with technology, the chapter provided information about some m-business software solutions.

Future of M-Business

With the arrival of 3G wireless networks in the market and development of powerful mobile devices, shift in business, and need for more ubiquity, in few years we will be participating in and watching TV-quality business meetings of your corporation on your laptop while in the air (flying back home), coordinating and managing your business while enjoying the beach, evaluating and managing a project and assigning new tasks while interacting with nature, processing loan applications while on the road, and conducting other serious business activities on the streets. What is most amazing is that neither a manager, nor employees or colleagues will realize that all these times you were miles away from the office, because m-business will provide you with facilities to work as if in the office all the time.

In short, you will be carrying your office or enterprise in your briefcase because you will be doing m-business. Your office and business will reside at your fingertips, and you will not be tied to your office. Then, we also may not need to maintain large office buildings. And all this will be in the near future because m-business is the business of the future. At present, however,

there are more than enough challenges and issues in adapting m-business and transforming the traditional way of doing business. We will see its full realization in the next few years.

REFERENCES

- Deitel, H., Deitel, P., & Steinbuhler, K. (2001). *E-business and e-commerce for managers*. Upper Saddle River, NJ: Prentice-Hall.
- Deitel, H., Deitel, P., Nieto, T., & Steinbuhler, K. (2003). *The complete wireless Internet and mobile business programming training course*. Englewood Cliffs, NJ: Prentice-Hall.
- ICMB. (2005, July 11-13). *Proceedings of the 4th International Conference on Mobile Business*, Sydney, Australia. Retrieved from <http://www.mbusiness2005.org/>
- Kalakota, R. (2005). *Mobile business: Vision to value*. Retrieved from <http://www.kalakota.com/speakingtopics/mbusiness.htm>
- Kalakota, R., & Kurchina, P. (2004). *Mobilizing SAP: Business processes, ROI and best practices*. Mivar Press.
- Kalakota, R., & Robinson, M. (2001). *M-business: The race to mobility*. New York: McGraw-Hill.
- Nokia. (2005). Retrieved March 2005 from <http://www.nokia.com/mobilebusiness/americas/>
- Paavilainen, J. (2002). *Mobile business strategies*. Reading, MA: Addison-Wesley.
- Sadeh, N. (2002). *M-commerce: Technologies, services, and business models* (1st ed.). New York: John Wiley & Sons.
- SAP. (2005). Retrieved March 2005 from <http://www.sap.com/solutions/mobilebusiness/index.epx>
- Segan, S. (2005). Wireless without borders: Networks for those on the go. *PC Magazine*, 24(5), 90.
- Solheim, S. (2005). 3G gets real. *eWeek*, 22(3), 28.
- Vos, I., & de Klein, P. (2002). *The essential guide to mobile business*. Upper Saddle River, NJ: Prentice-Hall.

Section XI

Case Study

Chapter LII

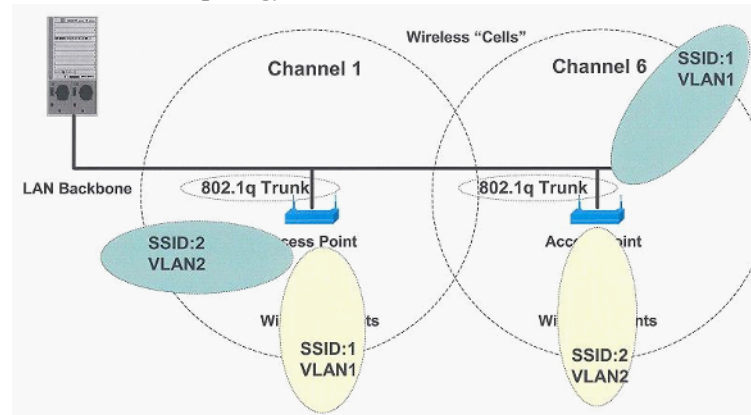
Successful Implementation of Emerging Communication Technologies in a Mobile-Intense Organization: A Case Study of Sydney Airport

Keyurkumar J. Patel
Box Hill Institute, Australia

ABSTRACT

Wireless Technology is growing at a phenomenal rate. Of the many present challenges highlighted by the author, increased security is one of the main challenges for both developers and end users. This chapter presents this important security aspect of implementing a mobile solution in the context of Sydney International airport. After tackling initial challenges and issues faced during the implementation of wireless technology, this chapter demonstrates how security issues and wireless application were implemented at this mobile-intense airport organization. The decision to deploy and manage the wireless spectrum throughout the Airport campus meant that the wireless LAN had to share the medium with public users, tenants and aircraft communications on the same bandwidth. Therefore, this case study also demonstrates invaluable approach to protect unintended users from breach of existing security policies adopted by their corporate network. Authentication and data privacy challenges, as well as complete WLAN connectivity for tenants, public and corporate usage is presented in this case study.

Figure 1. Typical wireless LAN topology



INTRODUCTION

Sydney's International Airport forms the hub of aviation in the Pacific region. It is an 85-year-old site, approximately 8 kilometers from Sydney CBD. With 5 terminals with 34 international, 31 domestic, and 5 airfreight gates, it is the largest airport catering to 8.7 million international and 15.5 million domestic passengers per year (McCubben, 2003). As such, an acute need was felt to ensure a high level of timely and quality service to the entire infrastructure of the airport. Mobile technologies were considered as a crucial ingredient in provision of this service. This need continues to be corroborated worldwide; for example, at the Airport Council International (ACI) World Assembly in Santiago in November 2000, the airport community expressed the importance of a wireless infrastructure at airports managed by the airport authority (Sydney Airport Corporation Limited, 2000). The following resolution was agreed upon:

Airport Operators should assert control over the use of Wireless Infrastructure at Airports, both inside and outside terminal buildings. Tenants, concessionaires and others should use a common infrastructure for wireless managed by the Airport Operator. In return

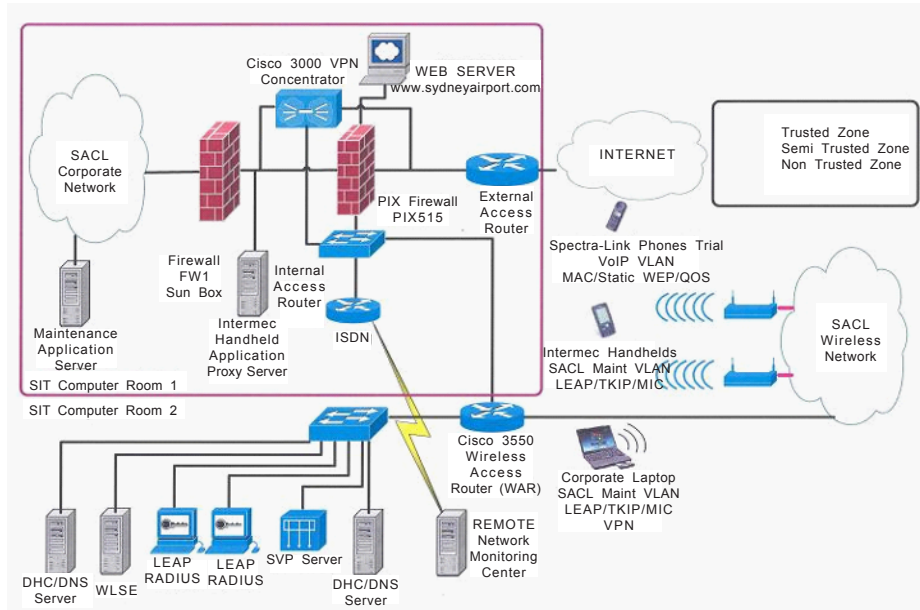
for this exclusivity, Airport operators should constantly evaluate competing technologies, so as to maintain low costs, increased capacity and security in line with demand for the benefits of all tenants, concessionaires and others.

Meanwhile, in 1998, with the impending Sydney Olympics 2000, Sydney Airport Corporation Limited (SACL) was formed. SACL took it upon itself to embark on the challenge of becoming the sole provider of wireless infrastructure at the International Terminal and Airfield. Past experience indicated that business customers preferred to install their own networks, and wireless—still an evolving technology with no ratified security standards and ease of deployment—gave SACL a unique challenge. This chapter discusses in detail the successful deployment of mobile applications at the Sydney International Airport.

WLAN Architecture and Security Challenges

With a typical wireless LAN (WLAN; see Figure 1), transmitted data is broadcast over the air using radio waves. With a WLAN, the boundary for SACL's network has moved and is now located in many airfield remote sites. In

Figure 2. Network architecture and security policy



early 2001, SACL deployed some 120 access points within the International Terminal (Terminal 1) and at various sites on the airfield. Sydney Airport WLAN implementation in early 2001 deployed Cisco Aironet 350 Series Access Points. The IEEE 802.11b standard adopted uses the unlicensed 2.4x gigahertz frequency band, providing only three non-overlapping channels (1, 6, and 11) with data-rates of 1, 2, 5.5, and 11Mbps.

However, without stringent security measures in place, the wireless infrastructure is equivalent of putting Ethernet ports everywhere. Thus, SACL's wireless deployment challenge was to ensure that the implementation of the wireless network did not breach its existing security policies for the corporate network. SACL regards the wireless network infrastructure in much the same fashion as the Internet, an untrusted zone. Even with this view, SACL has still ensured that wireless network security protects Sydney Airport's wireless VLAN. The following outlines the type of wireless and network security utilized at Sydney Airport.

Network Architecture and Security Policy

An important decision when deploying a WLAN is how it will interface back into the corporate infrastructure. The Wireless LAN (WLAN) at Sydney Airport has been designed so that the WLAN infrastructure access is located outside the corporate firewalls (see Figure 2). This approach creates more administrative overhead, because of the need for configuration of the External access network, consisting of router access lists and firewall rules. Furthermore, Sydney Airport does not need to maintain multiple WAN (Wide Area Network) remote sites. This is due to the fact that SACL's WLAN network is not deployed in order to replace the wired LAN in the office and is a network not solely accessed by SACL users. All external access from the wireless network is via the Wireless Access Router (WAR). The WAR is a Cisco 3550 Router configured to perform access-list filtering on all traffic based on source and destination IP address, protocol, and port numbers (Cisco Systems, 2002).

Many regard wireless technology as insecure (Arbaugh, 2001). SACL regards the wireless infrastructure at Sydney Airport as an untrusted zone. There are only two ways that a SACL wireless client can gain access to data from SACL's corporate network. The first option is to install an application proxy server. The proxy server allows data to move from the wireless untrusted zone to a semi-trusted zone located in the De-Militarized Zone (DMZ) outside SACL's Corporate Network, Firewall 1. This application proxy server located in the semi-trusted zone accesses the corporate network or trusted zone via Firewall 1 on the clients' behalf. The second option and only way resources can be accessed directly from an untrusted to a trusted zone is via the use of a Virtual Private Network (VPN). A VPN is established between SACL's wireless client and the Cisco VPN concentrator connected in parallel to the PIX firewall located in the DMZ. Both of these methods are described in detail further on.

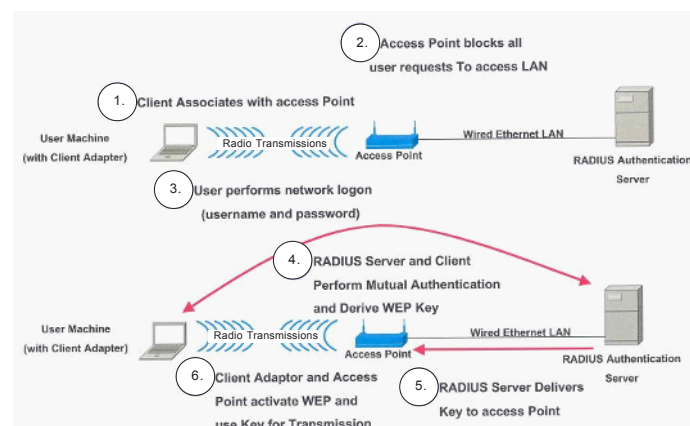
SACL Wireless Authentication

Originally, SACL only offered a Cisco Proprietary Wireless Security solution of Light-weight Extensible Authentication Protocol (LEAP) for

all tenants and concessionaires utilizing SACL's infrastructure. Cisco's LEAP utilizes a 128-bit dynamic Wireless Equivalent Privacy (WEP) key, along with radius username and password authentication (Geier, 2002). The WEP key is dynamically assigned from one of two Remote Authentication Dial-In User Service (RADIUS) servers (Cisco Secure ACS 2.6 or greater) when an authenticated user associates with an access point. This WEP key is again negotiated between the client and the RADIUS server after a pre-configured period set on the RADIUS server. At present this period is set at less than 10 minutes (see Figure 3).

The new firmware software introduced by Cisco for their access points supports the termination of 802.1q trunk. This allows a trunk to be provisioned between the access point and an Ethernet switch, the end result allowing users in a wireless VLAN cell to belong to different VLANs. With the use of different VLANs, user traffic is segmented per group (i.e., per VLAN) with the use of differentiated security policy per VLAN. The Service Set Identifier (SSID) is used to map the client to the wireless VLAN. RADIUS attributes passed in between the access point can also override this mapping if the users are not authorized for that SSID. Up

Figure 3. SACL wireless authentication—LEAP



to 16 wireless VLAN's can be supported on each access point.

The introduction of wireless VLAN's allows the use of non-priority client cards, along with different security models. Although SACL believes that the Cisco LEAP solution is still the most secure and manageable solution presently available, when used in conjunction with Temporal Key Integrity Protocol (TKIP) and MIC, (TKIP and MIC security are explained in the next section), it is possible for a static WEP, Media Access Control (MAC), and or EAP security options to be used on a separate wireless VLAN. This enables the use of all vendors' 802.11b wireless clients adaptor and gives greater flexibility for products not yet supporting LEAP, such as most voice over IP wireless phones. At present SACL continues to use the LEAP solution with TKIP and MIC, as all their devices make use of Cisco client adapters.

Wireless Data Privacy Enhancements

While WLAN security that relies on Service Set Identifiers (SSIDs), open or shared-keys, static WEP keys, or MAC authentication is better than no security at all, it is not sufficient or truly manageable for the size of the Sydney Airport wireless network.

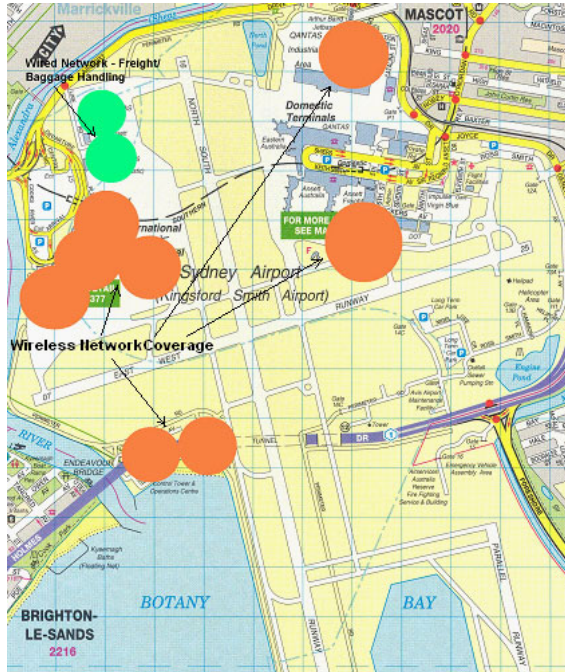
Sydney Airport, like all wireless network administrators, eagerly awaits the wireless IEEE 802.11i security standard that will allow vendor interoperability and still solve all known vulnerabilities of WEP (Stubblefield, 2002), the basic mechanism to date for interoperable security of Wireless 802.11b products. The IEEE 802.11i standards were published at the end of 2003. The Wi-Fi Alliance represented by many of the wireless vendors and in conjunction with the IEEE, has driven an effort to bring strongly enhanced, interoperable Wi-Fi security to market in the first quarter of 2004. The result of this effort is Wi-Fi Protected Access (WPA). Wi-Fi Protected Access is a specification of stan-

dards-based, interoperable security enhancements that strongly increase the level of data protection and access control for existing and future wireless LAN systems. Designed to run on existing hardware as a software upgrade, WPA is derived from and will be forward compatible with the upcoming IEEE 802.11i standard. One 802.11i component not required in WPA is Advanced Encryption Standard (AES) support. AES will replace 802.11's WEP initialization Vector RC4-based encryption under 802.11i specifications. Migrating to AES encryption, though, will require hardware changes, so this has been deferred by the Wi-Fi Alliance until the formal standard is in place to give vendors and customers some breathing room. The bad news is that 802.11i will require hardware changes regardless of whether WPA gets deployed over the next year or not.

Cisco has already been given WPA certification on the new IOS software available for both the 1100 and 1200 series Access Point range, with firmware for the 350 Series Access Points installed at Sydney Airport scheduled for the third quarter of 2005. Until the release Sydney Airport utilized Cisco's proprietary solution that features a subset of the 802.11i draft. As mentioned previously, Cisco has developed an 802.1X authentication type called EAP Cisco Wireless, or Cisco LEAP. Access points at Sydney Airport can be configured to support Cisco LEAP and all 802.1X authentication types, including EAP Transport Layer Security (EAP-TLS provides for certificate-based mutual authentication that relies on client-side and server-side digital certificates). With 802.1X authentication types such as LEAP and EAP-TLS, mutual authentication is implemented between the client and a RADIUS server. The credentials used for authentication, such as a log-on password, are never transmitted in the clear, or without encryption, over the wireless medium. Another benefit of 802.1X authentication is centralized management of WEP keys. Once

Successful Implementation of Emerging Communication Technologies

Figure 4. SACL wireless deployment (map © Sydway Publishing Pty. Ltd. Reproduced with permission)



mutual authentication has been successfully completed, the client and RADIUS server each derive the same WEP key, which will be used to encrypt all data exchanged. The result is per-user, per-session WEP keys. AP software running at the airport provides several enhancements to WEP keys that have formed part of the Wi-Fi Protected Access. These WEP enhancements include Cisco's pre-standard Temporal Key Integrity Protocol and support for Message Integrity Check (MIC).

When TKIP, also known as key-hashing support, is implemented on both the AP and all associated client devices, the transmitter of data hashes the base key with the IV (Initialization Vector of RC4 Key Scheduling Algorithm) to create a new key for each packet (Fluhrer, 2001). By ensuring that every packet is encrypted with a different key, key hashing re-

moves the predictability that an eavesdropper relies on to determine the WEP key by exploiting IVs.

When MIC support is implemented on both the AP and all associated client devices, the transmitter of a packet adds a few bytes (the MIC) to the packet before encrypting and transmitting it. Upon receiving the packet, the recipient decrypts it and checks the MIC. If the MIC in the frame matches the calculated value (derived from the MIC function), the recipient accepts the packet; otherwise, the recipient discards the packet. Using MIC, packets that have been (maliciously) modified in transit are dropped. Attackers cannot use bit-flipping or active replay attacks to fool the network into authenticating them, because the MIC-enabled client and access points identify and reject altered packets.

Mobile Maintenance

The initial deployment of the wireless network within Terminal 1 and the airfield coincided with the Mobile Maintenance Project. The project utilized the wireless network to track and complete maintenance work in the field. Maintenance staff at Sydney Airport uses a Computerized Maintenance Management System (CMMS) known as MAXIMO (see Figure 5). The one limitation of MAXIMO was its inability to follow staff to the job. Previously "work orders" or job sheets were printed directly off the system and then taken into the field by the relevant trade staff or technician. Once the work was completed, the sheet was completed manually and the written data then entered in MAXIMO, either on return to the workshop or at the end of a shift. The mobile solution utilizing an industrialized handheld personal digital assistant (PDA) eliminates re-entry of data and has a positive follow-on effect in allowing more accurate reporting and a paperless system. With the limitation of VPN

Figure 5. Computerized maintenance management system

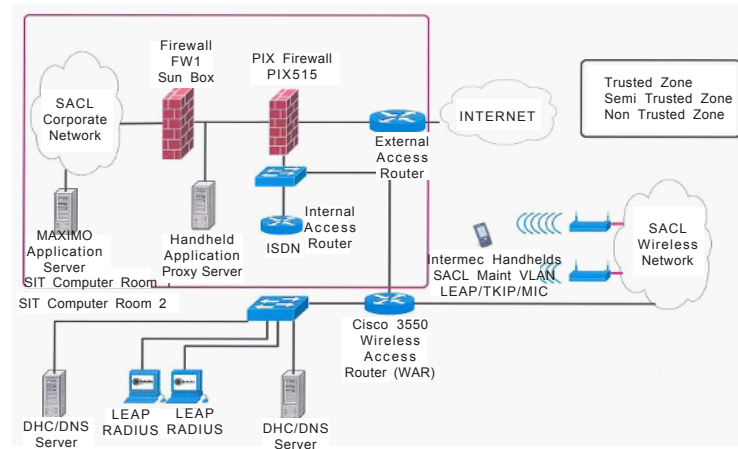
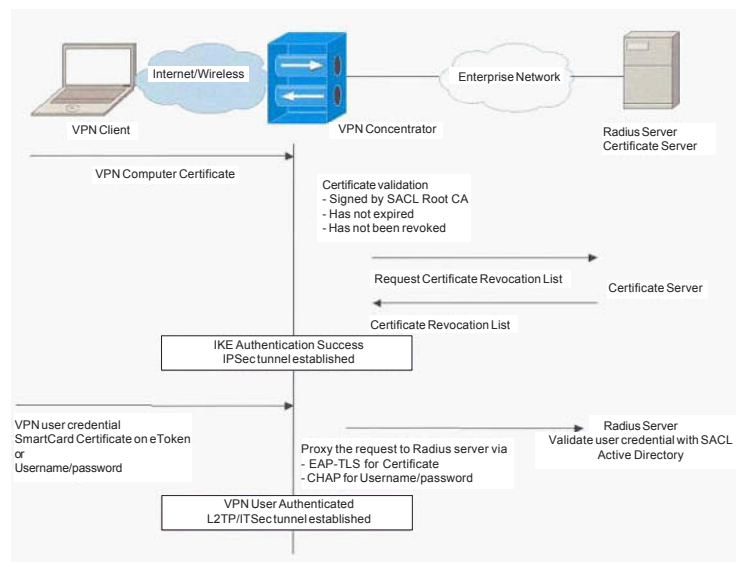


Figure 6. VPN authentication process



clients for PDAs at the time of delivery for the project, an application proxy server was installed within SACL's semi-trusted DMZ area to proxy wireless client requests to the MAXIMO application server. The PDA used was fitted with a Cisco Client Adaptor card and set up to use LEAP. The only changes from the initial installation are the upgrade of wireless client driver to utilize Cisco's TKIP and MIC

security enhancements configured on the access points.

Wireless VPN Remote Access Solution

In December 2002 SACL embarked on a corporate remote access solution. The solution enables SACL corporate users to access the

corporate network with all desktop applications from broadband Internet access, corporate dialup, and wireless access. By placing the wireless infrastructure access outside the corporate firewalls, it allows SACL to best utilize its remote access VPN solution. The implemented solution integrates a Cisco VPN 3000 Concentrator, Microsoft 2000 Certificate Server, Microsoft VPN Client, USB Port Token, and Centralized remote PC firewall (ZoneLabs Integrity) to provide strong security, ease of use, and centralized management. The original goal was to utilize the Cisco VPN Client that allowed cooperative reinforcement with the remote PC firewall. The ZoneLabs VPN Enforcement feature ensures that the VPN users can only connect to and remain connected to the SACL network as long as the client is running a verified version of the ZoneLabs firewall agent and the client is enforcing the most up-to-date security policy.

The Cisco VPN client used was incompatible with a few of the corporate applications which required the client to have a virtual adaptor IP address in order to ftp data from a server back to the remote PC. Cisco released a VPN client that has a virtual adaptor which is still being piloted by SACL. In the interim period, SACL chose to utilize a Microsoft VPN client that does not provide cooperative reinforcement of the remote PCs. Remote PCs have utilized the Microsoft XP operating system to lock down the ZoneLabs application to ensure that the user cannot shutdown the firewall. As all the wireless users within SACL are utilizing Cisco wireless adaptors, the same Wireless VLAN utilizing LEAP is used. This will change in the near future, as the new laptops have inbuilt wireless adapters. As the VPN solution does not require the additional wireless security, as it is geared for broadband and hotspot users, another wireless VLAN will be created with a different security model. Figure 2 shows the present VPN establishment

and logon process to the corporate network. Strong authentication is required to secure the VPN connections (see Figure 6). VPN users must have a computer with a valid SACL digital certificate, valid Windows account, or USB eToken with SmartUser certification to successfully establish a VPN connection. After the use of valid Digital Certificate for Internet Key Exchange (IKE) authentication, the Microsoft VPN uses Internet Protocol Security (IPSec) Encryption ESP (encapsulation protocol)—Layer Two Tunneling Protocol (L2TP)—Transport IPSec SA.

Wireless Voice Over IP Pilot

The third device that has been trialed by SACL across its wireless LAN is a pilot of wireless IP phones. During the last quarter of 2002 SACL, undertook a voice over IP trial. This pilot included the deployment of wireless handsets that ran over the terminal's live wireless network. The wireless IP handsets used were spectra-link phones. These phones did not support LEAP, so a dedicated WLAN with static 128-bit WEP was set up across the international terminal. This wireless VLAN was also given the highest quality of service (QoS) on each access point to ensure phone calls would not drop out if a wireless access point was supporting multiple clients. Knowing the security vulnerabilities of static WEP, SACL combined MAC-level security on the wireless VLAN. MAC-level security on the Cisco access points can be centrally managed by the Cisco Secure ACS RADIUS Servers. This is performed by entering the MAC address of the phone as a user and password in the RADIUS server. While the security solution is not necessarily ideal, it was the only means available at the time of the pilot. Cisco's latest release of software for the 1100 and 1200 access points has a feature known as fast secure roaming.

This feature allows EAP authentication to be used for Wireless Voice Over IP.

Tenant Wireless Connectivity

As the wireless infrastructure is a shared medium for both SACL and its tenants, it is necessary to establish connections back into the tenants' own corporate network. Figure 8 shows how this connection is implemented, indicating demarcation points. A range of IP addresses is assigned to each tenant who will either hard code or utilize SACL's DHCP servers. Every tenant utilizing SACL's wireless infrastructure will be given a dedicated 100 Mbps UTP routed connection to the Wireless Access Router. Within the WAR, access-lists filtering on IP address, protocol, and port numbers are configured. It is advised that the 100 Mbps Ethernet port be connected to the tenants' network via their own firewall. The tenant is responsible for any additional security measures such as VPNs with their own wireless clients. The tenant is set up with their own Wireless VLAN that can be given a security policy that best meets their needs or device capabilities.

Until, the introduction of Wireless VLANs, there was only one option—to use Cisco client

adaptors utilizing LEAP. This interoperability proved to be challenging, with one of the International Airline Lounges already entered into a commercial agreement with an ISP to provide wireless Internet access to their Frequent Flyer Members. As SACL could not control the public users' client adaptor card, an interim solution was put in place. Other tenants wanting to utilize symbol barcode scanning devices were told the warranty on the device would be void if they replaced the wireless adaptor with a Cisco Card. With the release of wireless VLANs, two organizations utilizing SACL's wireless network can now use symbol devices with symbol wireless client adaptors. Both are using static WEP with Symbols own VPN solution AirBEAM Safe.

Public Internet Connectivity

Sydney Airport deployed a public wireless network (see Figure 7) to allow high-speed Internet connectivity to public users. SACL is strategically well placed to target this market of mobile professionals. SACL has partnered with Internet service providers to offer their existing customers and new subscribers all the Internet ser-

Figure 7. WLAN public Internet connectivity

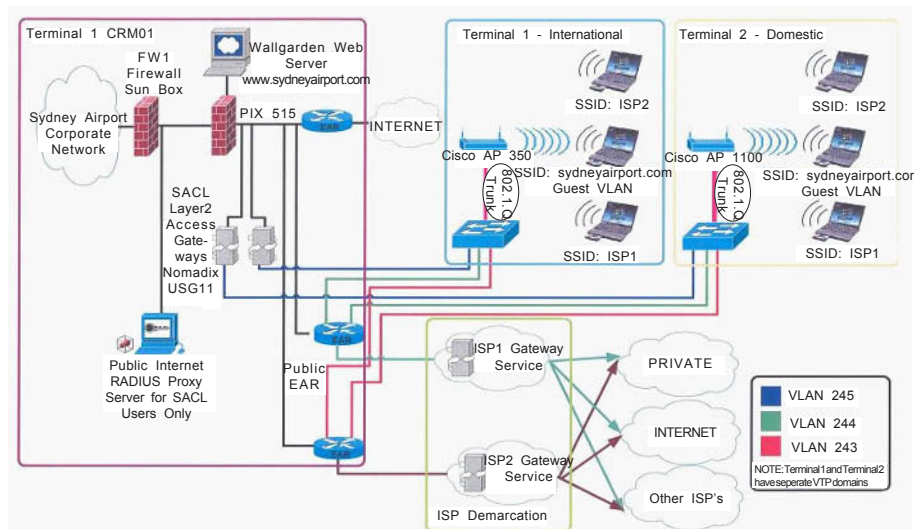
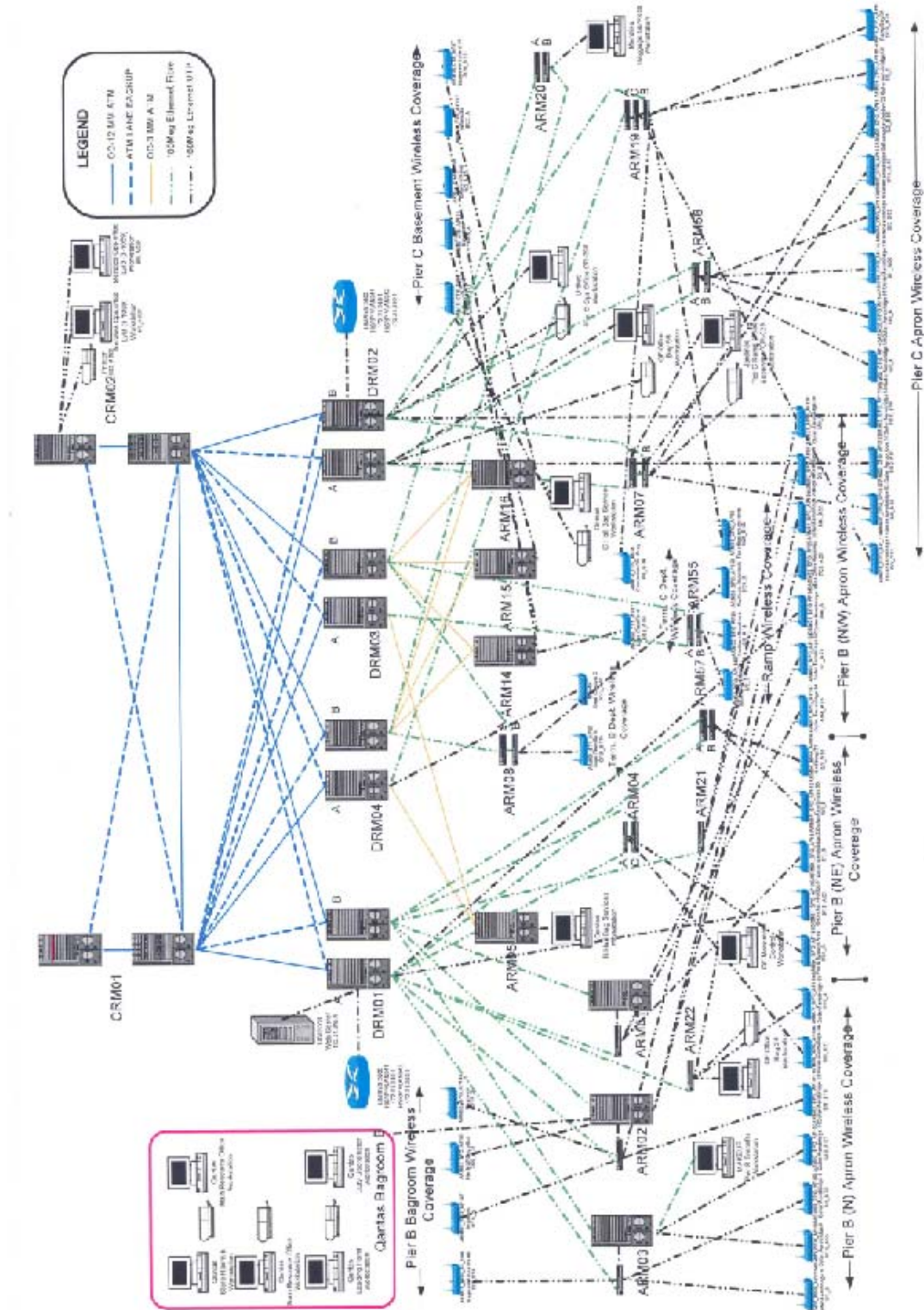


Figure 8. Total wireless connectivity diagram



vices including e-mail, Web browsing, and connection back into their corporate networks via secure VPN. Like most hot spots, the wireless public LANs are set to open authentication with no WEP key encryption configured. The portal page login for the subscriber is made by opening an SSL-encrypted session hypertext transfer protocol over Secure Socket Layer (HTTPS). The difference between most hotspots is that Sydney Airport hosts three unencrypted wireless LANs on each of its access points. As Sydney Airport already had the wireless infrastructure installed, it made commercial sense to cater for multiple service providers instead of the usual “ISP-grabbing-real-estate” approach. The commercial model with each of the ISPs is based on revenue sharing of subscriber usage. The subscriber can be classified as retail and wholesale. A wholesale customer refers to a subscriber who is signed with a roaming partner of the ISP. To cater for multiple ISPs proved quite challenging. The technical challenges of routing users down different Internet service connections, government regulations on IP interception rules in Australia, and the airport not wanting to lawfully become an Internet service provider led to the decision of running multiple wireless VLANs. The latest 350 IOS firmware for Cisco access points allows for multiple unencrypted VLANs. SACL therefore maintains control of the access point infrastructure and WLAN services on the airport site, being responsible for ensuring the WLAN technical capability of the network. Each ISP is provided with a non-broadcast unencrypted wireless LAN, with their nominated Wireless Network Name (SSID) as set up in their other hotspot sites. This WLAN will be connected to the ISP’s gateway service infrastructure. The Internet service provider will ensure compliance with the relevant government regulations. This compliance will mean that they will solely be responsible for interception of all IP traffic pertaining to their WLAN on the airport site.

The Internet service provider shall provide to SACL a non-repudiable system for reporting and auditing of the Sydney Airport hotspot site. The system shall be accessible to SACL in real time or close to it. Sydney Airport requires online data such as billing or RADIUS accounting records that detail each user’s IP session time along with data byte usage.

CONCLUSION AND FUTURE DIRECTION

The security interoperability challenges are being addressed by Wi-Fi Protected Access range of Wi-Fi products based on the upcoming IEEE802.11i standard. With the WPA and ultimately 802.11i standard implementation in place, the need for add-on solutions such as VPNs may be deemed unnecessary in some enterprise wireless LAN environments (Patel & McCubben, 2004a). Other future enhancements not referenced in this chapter include the upcoming release of 802.11g Wireless Access Points. 802.11g can ideally deliver 54 Mbps maximum data rate and offer an additional and compelling advantage—backward compatibility with 802.11b equipment. This means that 802.11b client cards will work with 802.11g access points, and 802.11g client cards will work with 802.11b access points. Because 802.11g and 802.11b operate in the same 2.4 GHz unlicensed band, migrating to 802.11g will be an affordable choice for Sydney Airport with existing 802.11b wireless infrastructures (see Figure 8).

One drawback is that 802.11b products cannot be “software upgraded” to 802.11g because 802.11g radios will use a different chipset than 802.11b in order to deliver the higher data rate (Patel & McCubben, 2004b). However, like Ethernet and Fast Ethernet, 802.11g products can be co-mingled with 802.11b products in the same network. Sydney Airport will con-

tinue to provide solutions to business needs by utilizing innovative and leading-edge technology. Future applications include Mobile Self-Service, Check-In Kiosks, RF Bag Tags, Wireless Point of Sale, Wireless Stock Take, Wireless VoIP, and the end-of-the-year government requirements for Baggage Reconciliation. To-date total wireless connectivity diagrams are shown in Figures 4 and 8. So far it is safe to say that the wireless technology will play an important part in the future of Sydney Airport's total journey experience.

REFERENCES

- Arbaugh, W. A., Shankar, N., Justin Wan, Y.C. (2001). *Your 802.11 wireless network has no clothes*. Retrieved June 20, 2003, from <http://www.cs.umd.edu/~waa/wireless.pdf>
- Cisco Systems. (2002). Retrieved June 20, 2002, from http://www.cisco.com/en/US/netsol/ns340/ns394/ns348/ns337/networking_solutions_package.html
- Fluhrer, S. R., Mantin, I., & Shamir, A. (2001). *Weaknesses in the key scheduling algorithm for RC4*. Retrieved November 11, 2001, from http://downloads.securityfocus.com/library/rc4_ksaproc.pdf
- Geier, J. (2002). *802.11 security beyond WEP*. Retrieved July 12, 2002, from <http://www.wi-fiplanet.com/tutorials/article.php/1377171>
- McCubben, S. (2003). *Trim document reference: M2003/06745*. Sydney: Sydney Airport Corporation Limited.
- Patel, K. J., & McCubben, S. (2004a). Addressing wireless security issues during implementation of wireless applications in a highly mobile organization. In *Proceedings of the International Conference on Computing, Communications and Control Technologies* (Vol. 7, pp. 13-18).
- Patel, K. J., & McCubben, S. (2004b). Implementation of wireless technology in a highly mobile organization: Challenges and issues. In *Proceedings of the 8th World Multi-Conference on Systemics, Cybernetics and Informatics* (Vol. 11, pp. 43-48).
- Stubblefield, A., Ioannidis, D., & Rubin, A. (2001). *Fluhrer, Mantin, Shamir attack to break WEP*. Retrieved June 20, 2002, from <http://www.isoc.org/isoc/conferences/ndss/02/proceedings/papers/stubbl.pdf>
- Sydney Airport Corporation Limited. (2000). Retrieved June 24, 2000, from <http://www.sydneyairport.com>

Chapter LIII

The Next Big RFID Application: Correctly Steering Two Billion Bags a Year Through Today's Less-Than-Friendly Skies

David C. Wyld

Southeastern Louisiana University, USA

ABSTRACT

This chapter examines the adoption of radio frequency identification (RFID) technology in the commercial aviation industry, focusing on the role of RFID systems for improved baggage handling and security. The chapter provides a timely overview of developments with regard to the implementation of RFID technology in commercial aviation, which promises distinct advantages over the currently used bar-code system for baggage handling. The chapter focuses on how RFID technology can improve customer service through better operational efficiency in baggage handling, which has been demonstrated to be an integral component of the airline's customer service equation. Developments with RFID technology can dramatically improve the accuracy of baggage handling, which can enable air carriers to close an important service gap among customers in an increasingly turbulent operating environment. Other service industries can certainly benchmark the airline industry's use of RFID technology in luggage tracking as a way to improve their own operational capabilities.

INTRODUCTION

To put this chapter in perspective, consider this scenario: You have just landed in Alexandria, Egypt, or Alexandria, Louisiana. You are standing at the baggage carousel, having flown in on the last flight arriving that night. A constant stream of bags of all shapes, sizes, and colors circle past you, disappearing one by one as your

“lucky” fellow passengers claim their prizes. After about 15 minutes, the carousel stops spinning. At that point, you realize that your checked roller-bag has not arrived on the same flight as you.

Now, you are in “lost luggage hell,” and while the airline may do its best to accommodate you, no amount of compensation from the air carrier—whether in money, miles, or drink

coupons—can change one simple fact: How are you going to make that winning presentation to a major new client at 8:00 the next morning? You realize that the only clothing you have in your possession is the warm-up suit you wore to be comfortable all day as you traveled; your “killer suit” and “confidence tie” are likely sitting on an airport tarmac thousands of miles away, with no clothing store in the city that will open before the meeting (unless you happen to be in Las Vegas).

The system that you are dependent upon to correctly track your checked luggage to either the Memphis in Tennessee or in Egypt, or wherever else it may be, is based on correct readings along the line of a bar-coded label, bearing a 10-digit IATA (International Air Transport Association) number. Gartner’s Research Director, Jeff Woods, commented that “bags are very well tracked right now” by the airlines and their bar code-based systems (cited in Morphy, 2004). Yet, this is little consolation when it is *your* bag that is lost. The baggage tracking systems of the world’s airlines are mature, and even under the best of conditions, bar code technology works in correctly reading only eight or nine bags out of every 10. This means that the airlines continue to devote considerable time and energy to manually intervene to correctly direct the right bags onto the right flights, while spending great amounts of money to reunite passengers with their bags when the system breaks down.

Today, savvy airlines, even in their precarious financial positions, are seeing the shift to RFID (radio frequency identification)-based baggage tracking systems as a solid operational investment that can produce significant cost savings and demonstrated return on investment (ROI). Airports as well are taking the initiative to shift to RFID-based systems, sensing the opportunity to produce greater traveler satisfaction with their experience at a specific airport. In a deregulated world of airline and

airport choices, these entities are combining forces to enhance customer service and give them a competitive advantage, perhaps for a significant window of time until such RFID-based systems are made mandatory.

In this chapter, we will examine the mechanics of how RFID-based baggage tracking works and the benefits it can provide. After a brief overview of RFID technology, we will look at the experience of Delta Air Lines, which is the first airline to publicly commit to taking the technological leap forward to implementing RFID-based baggage tracking. We will then examine the confluence of technology, terrorism, and yes, marketing, that will likely drive the adoption of RFID-based tracking of checked baggage throughout the world. The RFID movement is also being spearheaded by the U.S. government. It is clearly interested in securing the safety of the traveling public and with it, what financial viability the airline industry has left in the wake of the after-effects of September 11, 2001, and the decline in travel spurred by that awful tragedy, an economic recession, and record fuel prices. We will examine the government push in this area and concerns over passenger privacy. Finally, we will look at an alternative vision of the future of airline customer service, which may preclude the need for baggage service as part of the air passenger experience altogether.

WHAT IS RFID?

In brief, radio frequency identification uses a semiconductor (microchip) in a tag or label to store data. Data is transmitted from, or written to the tag or label when it is exposed to radio waves of the correct frequency and with the correct communications protocols from an RFID reader. Tags can be either *active* (using a battery to broadcast a locating signal) or *passive* (using power from the RFID reader for

location). A firm may use a combination of fixed and handheld readers for reading RFID tags to gain as complete a picture as has ever been possible on exactly what is where in their operations. Reading and writing distances range up to 100 feet, and tags can be read at high speeds (Booth-Thomas, 2003). For a detailed explanation of the technology, see Jones and Wyld (in press), McFarlane (2002), Kambil and Brooks (2002), and Reed Special Supplement (2004).

The advantages of RFID over bar code technology are summarized in Table 1.

RFID tags have been described as being a “quantum leap” over bar codes. *Inc. Magazine* characterized RFID vs. bar codes as “like going from the telegraph to the Internet” (Valentine, 2003).

As noted in an interview last year with the *Harvard Business Review*, William Copacino, group chief executive officer for Accenture’s Business Consulting Capability Group, interest in RFID is picking up significantly throughout the global business community today. This is due not only to the fact that prices are rapidly dropping for both the RFID tags themselves and for the readers to sense them, but more importantly, the technology is providing significant improvements in operations and efficiency

over traditional methods, while affording companies the concomitant opportunity to improve their customer service strategies (opinion cited in Kirby, 2003). From the perspective of Deloitte Consulting (2004), if RFID is viewed as simply an alternative means of identification and labeling to bar code technology, then businesses will have a “lost opportunity” on their hands. This is because RFID technology potentially offers wide-ranging opportunities for transformative change (a change of the highest magnitude) in internal business processes, supply chain management, security threat management, and customer service. Innovative applications of RFID technology are being seen in myriad industries today, including such critical areas as pharmaceuticals (Wyld & Jones, in press) and livestock tracking (Wyld, Juban, & Jones, 2005).

BAGGAGE AND AIRLINE CUSTOMER SERVICE

The critical link in customers’ minds between seeing their luggage on the baggage carousel upon arrival and their perception of the quality of an airline’s service offering has been empirically proven. Each year, professors Brent D. Bowen (University of Nebraska Omaha) and

Table 1. RFID vs. bar code technology

Bar Codes	RFID Tags
Bar codes require line of sight to be read.	RFID tags can be read or updated without line of sight.
Bar codes can only be read individually.	Multiple RFID tags can be read simultaneously.
Bar codes cannot be read if they become dirty or damaged.	RFID tags are able to cope with harsh and dirty environments.
Bar codes must be visible to be logged.	RFID tags are ultra thin, and they can be read even when concealed within an item.
Bar codes can only identify the type of item.	RFID tags can identify a specific item.
Bar code information cannot be updated.	Electronic information can be overwritten repeatedly on RFID tags.
Bar codes must be manually tracked for item identification, making human error an issue.	RFID tags can be automatically tracked, eliminating human error.

Table 2. 2004 airline quality ratings (adapted from Bowen & Headley, 2004)

RANK	AIRLINE	AQR SCORE
1	JETBLUE AIRWAYS	-0.64
2	ALASKA AIRLINES	-0.74
3	SOUTHWEST AIRLINES	-0.89
4	AMERICA WEST AIRLINES	-0.89
5	US AIRWAYS	-0.96
6	NORTHWEST AIRLINES	-1.02
7	CONTINENTAL AIRLINES	-1.04
8	AIRTRAN AIRWAYS	-1.05
9	UNITED AIRLINES	-1.11
10	ATA AIRLINES	-1.17
11	AMERICAN AIRLINES	-1.24
12	DELTA AIR LINES	-1.24
13	AMERICAN EAGLE AIRLINES	-2.10
14	ATLANTIC SOUTHEAST AIRLINES	-5.76
	INDUSTRY AVERAGE	-1.14

Dean E. Headley (Wichita State University) produce their *Airline Quality Rating* report. These researchers' analytical methodology ranks airline performance in the United States, based on a weighted average of four key performance measures. These benchmarks have been validated as key in determining consumer perceptions of the quality of airline services. The four measures, drawn from data that the airlines are mandated to report to the U.S. Department of Transportation, include:

1. on-time arrivals,
2. mishandled baggage,
3. involuntary denied boardings, and
4. 12 areas of customer complaints.

Several airlines in the U.S. that have performed well in the quality survey, including Southwest, JetBlue, and Midwest Express, have touted their rankings in Bowen and Headley's (2004) report in their advertising campaigns.

Such has not been the case with Atlanta, Georgia-based Delta Air Lines. Based on the recently released *Airline Quality Rating 2004* report (as seen in Table 2), Delta has now fallen to *last* among the 12 major U.S. airlines in consumer perceptions of service quality. To put this in perspective, while the airline's composite quality rating has actually *improved* over

time since 2000, in that same timeframe, Delta's competition has been making marked improvements in the service components that matter most to airline customers.

Today, Delta is a firm embroiled in the turmoil that makes up the airline industry in America. Facing rising fuel costs, a downturn in business travel, an uncertain economy, and discount competition, *all* the established, legacy carriers in the U.S. are struggling financially and operationally today, with prominent carriers such as US Airways and United barely surviving (e.g., see Tully, 2004). Delta itself has been the subject of bankruptcy rumors, and it has conducted layoffs and closed its major hub at the Dallas/Fort Worth International Airport to stave off its demise (Perez, 2004). In September, CEO Gerald Grinstein announced a comprehensive overhaul plan, including laying off thousands of employees, and received initial agreement from its pilots' union to the recall of retired pilots on a limited basis (Fein, 2004; Weber, 2004). The airline industry is finding that without the ability to raise fares or to spend lavishly to improve customer service, it must improve its operational efficiencies and performance to survive today.

One particular area of weakness for Delta has been its handling of air travelers' checked-in luggage. In fact, according to the recently released 2004 report (which uses annual data as of the close of 2003), Delta's mishandled baggage rate increased from 3.57 in 2002 to 3.84 in 2003. As can be seen in Table 3, Delta still remains below the industry average rate of four lost bags per 1,000 passengers. However, Delta's own performance is impacted by that of Atlantic Southeast Airlines (ASA), Delta's regional partner throughout much of the United States. ASA "earned" *the* lowest quality rating of *all* airlines operating in the United States, regardless of size. Luggage service is a particularly sore point for Delta's code-sharing

partner, as ASA's rate of 15.41 mishandled bags per 1,000 passengers is almost *four times* the industry average.

Despite years of trying to improve the quality of its baggage-handling systems, Delta has seen the performance of its current bar code-based system flat-line, with bar-coded labels being successfully read by scanners only 85% of the time. According to Delta spokesman Reid Davis, the airline faced the fact that it had "reached the end of the improvements that could be accomplished without new technology" (cited in Rothfeder, 2004). Of course, just because a bag is not scanned correctly does not mean that your bag will end up in Wichita Falls when you were heading to Wichita. In the end, Delta estimates that only 0.7% of all checked luggage is actually "lost." However, the airline spends upwards of \$100 million each year to return these bags to their rightful owners and provide compensation to passengers whose luggage is never found (Collins, 2004).

Delta's top management has decided to tackle its "bag problem" head-on, looking to RFID technology as the means to an end of

providing far-better luggage service to its passengers. In the fall of 2003, Delta implemented a pilot test of an RFID tracking system for checked luggage on flights between Jacksonville, Florida, and its hub in Atlanta, Georgia. In this testing program, Delta tracked 40,000 passenger bags equipped with radio frequency identification (RFID) tags from check-in to loading on an aircraft. As can be seen in Table 4, the RFID-enabled system provided far superior reading accuracy than the legacy bar code-based system. In the spring of 2004, Delta implemented another pilot RFID baggage-tracking system at its Cincinnati, Ohio, hub, producing similar results (Murray, 2004).

Through the two test programs, Delta learned several valuable lessons. It saw that tag antennas could be damaged by the static electricity generated along the conveyor systems (Collins, 2004). It also found that the lowest scanner accuracy rate (96.7%) was found when attempting to scan bags inside the unit load devices (ULDs), the large containers pre-loaded with checked luggage that are then loaded onto the plane. The ULDs are made of metal with canvas doors, and the metal housing impeded the radio signals. Delta plans to coat the ULDs with a material that can better reflect the radio waves (Brewin, 2004a). While the test programs were conducted in rather neutral weather environs, concerns were raised over the ability of the tagging systems to function in harsher environments, such as at Delta's western hub in Salt Lake City, Utah (Murray, 2004). Finally, there is a famous American commercial from Samsonite that shows a gorilla in his cage, tossing the bag around and eventually stomping on the suitcase. The obvious message is that checked bags are not always handled "delicately" by the humans or the machinery as it passes through baggage systems. Thus, it must be noted that baggage handling itself can damage or detach labels/tags, and concerns over

Table 3. Mishandled baggage reports for U.S. airlines—June 2004 (adapted from U.S. Department of Transportation, Air Travel Consumer Report, August 2004; <http://airconsumer.ost.dot.gov/reports/2004/0408atcr.doc>)

RANK	AIRLINE	REPORTS PER 1,000 PASSENGERS
1	JETBLUE AIRWAYS	2.81
2	AIRTRAN AIRWAYS	3.02
3	SOUTHWEST AIRLINES	3.16
4	HAWAIIAN AIRLINES	3.18
5	ALASKA AIRLINES	3.32
6	CONTINENTAL AIRLINES	3.32
7	AMERICA WEST AIRLINES	3.55
8	NORTHWEST AIRLINES	3.80
9	ATA AIRLINES	3.80
10	UNITED AIRLINES	3.83
11	US AIRWAYS	4.10
12	DELTA AIR LINES	4.23
13	AMERICAN AIRLINES	4.66
14	EXPRESSJET AIRLINES	5.29
15	AMERICAN EAGLE AIRLINES	9.00
16	COMAIR	10.21
17	SKYWEST AIRLINES	10.71
18	ATLANTIC COAST AIRLINES	13.42
19	ATLANTIC SOUTHEAST AIRLINES	13.97

The Next Big RFID Application

Table 4. Results from Delta Air Lines pilot RFID test in Jacksonville (adapted from AIM Global, 2004)

Errors per 40,000 bags	RFID	Bar Code
Worst Case	1,320 (96.7%)	8,000 (80.0%)
Best Case	80 (99.8%)	6,000 (85.0%)

the durability of the RFID tag are genuine.

Even with limited capital to invest in IT projects, in July 2004 Delta became the first airline to commit to having RFID-enabled baggage tracking in place system-wide by 2007. Delta plans to use passive tags, which will cost the airline 25 cents each initially. However, the airline hopes that the cost of the tags will drop to approximately 5 cents a unit by the time the system is fully implemented in 2007 (McDougall, 2004). Delta estimates that the full implementation cost of its RFID-based tracking system will ultimately fall somewhere between \$15 and \$25 million for its 81 airport locations. Delta has not yet announced plans for deploying the RFID-based system with its code-sharing partners, which would greatly raise the number of airports worldwide for implementation and the cost and complexity of the overall project (Murray, 2004). While this represents a significant investment, the ROI equation shows that this cost can be recouped in far less than a single year. This makes Delta unique, as it is one of the few examples to date in *any* industry where the decision to invest heavily in automatic identification technology is based on the desire to dramatically improve customer service.

Delta's RFID-enabled baggage system will give the company the ability to track a bag from the time a passenger checks it in at his/her departing airport till the time the bag is claimed at the baggage carousel at the arrival airport. At check-in, the RFID tag's serial number will

be associated with the passenger's itinerary. Delta will position fixed readers at check-in counters and on conveyor belts where the bags are sorted. The airline will also equip baggage handlers with portable readers and outfit aircraft cargo holds with readers built into them. RFID readers can also be positioned to scan bags as they are loaded and unloaded from the unit load devices (ULD—the large containers that are loaded onto the plane). Through this surveillance system, Delta should be able to all but eliminate the problem of misloaded and misdirected checked luggage, and the attendant costs of reuniting the lost bag with the passenger. Ramp and flight crews will be able to make certain that the right luggage is on board before an aircraft takes off. And, in the event a passenger's bag is misdirected, Delta can instantly locate the bag through its RFID reader and more quickly route it to the passenger's destination. Pat Rary, a Delta bag systems manager, illustrated the fact that RFID will allow the airline to take proactive customer service steps on baggage problems. He observed that:

With this technology, we won't have to wait for the customer to come tell us that the bag is lost. We can tell the customer it's on the wrong plane and start responding before it's a crisis. Eventually, RFID should be able to signal an arriving passenger's cell phone with news of how long it will be before the bag is on the carousel. (cited in Field, 2004, p. 61)

Rob Maruster, Delta's director of airport strategy, recently commented in *Airline Business* that RFID tracking "will transform the airline on the ramp as much as radar did to transform air traffic control. When that happened, it was as if a light was turned on and people said, oh, so that's where the planes are.

This technology will do that for bags. People will say, oh, so that's where the bags are" (cited in Field, 2004, p. 60). Delta's ultimate goal is to have a baggage tracking system that will have a "zero mishandling rate" (Brewin, 2004a).

RFID AND BAGGAGE SECURITY

Unfortunately, in our post-September 11 world, there are worse things that can happen in the air or at the airport than losing one's luggage or even eating the "Chef's Surprise" at the airport restaurant. The twin, nearly simultaneous jet crashes in Russia in August 2004 have now been attributed to in-flight bomb detonations by Chechen female suicide bombers, raising fears that suicidal terrorists could use similar methods to attack the West (Hosenhall & Kuchment, 2004). While enhanced physical passenger screening, such as that just announced by the Transportation Security Administration (TSA) in the U.S., can deter such would-be suicidal terrorists, since September 11, the airline industry and national governments have placed renewed vigilance on screening both carry-on and checked bags for explosives and on making sure that all checked bags are matched to passengers who have actually boarded the aircraft. Writing in *Management Services*, Collins (2004) observed that one of the very real near-term applications for RFID technology is the prospect that a passenger's checked bag will be able to tell security personnel and the airline if it has not been properly screened.

The need for matching passengers with checked luggage has been at the forefront of anti-terrorism concerns ever since the in-flight bombings in the 1980s that took down a Pan American 747 over Lockerbie, Scotland, and an Air India jumbo jet over the Atlantic. Out of this concern, airlines must routinely remove bags from aircraft when a passenger fails to board,

out of fear that a homicidal, rather than suicidal, terrorist would attempt to down an airliner with a bomb in an unaccompanied, checked suitcase (AIM Global, 2004).

Often, this is a time-intensive, laborious process, which can delay flight departures indefinitely, as ramp workers face the daunting task of finding the bags in question out of the hold of an aircraft or from the unit load devices. Airport operations managers and airline flight crews will often tell horror stories of how the inability to find the one or two targeted bags of a non-boarding passenger in and amongst the bags of 300-400 passengers on a jumbo jet has caused flights to be delayed for hours, costing the airline countless amounts of goodwill amongst the passengers, even if such measures are done precisely to safeguard their transit and their very lives. Thus, airports are also very interested in providing better baggage tracking as part of their customer service equation.

In Florida, the Jacksonville International Airport installed an RFID-based system in 2003 to direct checked luggage through their newly installed baggage handling system. The city's airport authority and the TSA jointly funded the Jacksonville system. The contractors for the Jacksonville Airport project included FKI Logistex and SCS Corporation. The Jacksonville system was designed to only handle outbound luggage, directing checked bags from the check-in counter through explosive detection screening and on to the correct terminal serviced by the respective airlines. All checked bags have a bar code label affixed to them, with approximately 12% receiving an additional RFID tag, due to their being selected for special screening attention by a computer-assisted passenger profiling system (CAPPS) (Trebilcock, 2003).

The Jacksonville pilot program tested the effectiveness of both disposable and reusable tags. Passengers checking in on the north side

of the airport who were selected by the CAPPS had a disposable tag attached to their luggage, while those checking in on the airport's south side had a reusable, credit card size tag affixed to their checked bags. Each reusable tag costs \$2.40, and each disposable tag costs 63 cents. Van Dyke Walker, Jr., director of planning and development for the Jacksonville Airport Authority, believes that his airport's system is a precursor of what is to come. He commented that "RFID is the future of airline baggage tracking, and we want to be ready" (cited in Trebilcock, 2003, p. 40).

Las Vegas' McCarran International Airport is considered to be an ideal proving ground for RFID baggage tracking. This is due to the fact that the vast majority of the passengers using the airport either begin or end their journeys there. In fact, as Las Vegas sees only 8% of its passengers connecting to other flights at its airport, a rate that is only second to Los Angeles International Airport (Anonymous, 2003a), Las Vegas' system is designed to track all checked luggage, routing bags through bomb detection screening and on to the proper aircraft. From the perspective of Randall H. Walker, McCarran International Airport's director of aviation, the RFID-enabled baggage handling system "becomes a win for all concerned: the traveler, the airport, the TSA and the airlines" (cited in Anonymous, 2004a). In 2005, the TSA is slated to have similar systems in place at both LAX and Denver International as well (AIM Global, 2004). Alaska Airlines also uses the tags on its international flights out of San Francisco International Airport (Woods, 2004).

Internationally, RFID-based baggage tracking systems are being tested in Narita, Japan, Singapore, Hong Kong, and Amsterdam (CNETAsia, 2004; Atkinson, 2002). In fact, the RFID baggage tracking system being installed at Hong Kong International Airport is regarded

as the largest automatic identification system to be developed and deployed to date in Asia. Hong Kong's airport is one of the busiest in the world, handling approximately 35 million passengers each year. Y. F. Wong, who heads Technical Services and Procurement at the Airport Authority of Hong Kong, believes that the airport's investment in RFID technology is essential, as it addresses the need for improved customer satisfaction, while also enabling increased levels of security assurances (cited in Anonymous, 2004a). According to John Shoemaker, senior vice president of corporate development at Matrics, which will supply the airport with upwards of 80 million smart labels over the next five years, "What is key about Hong Kong International is that it is deploying this system to also save money" (quoted in Collins, 2004).

RFID baggage tracking is thus a means to an end for airports—with the end being improved baggage security. Simon Ellis, a supply-chain futurist at Unilever, recently observed that: "Security is just a sub benefit of visibility. Knowing exactly what is where gives you better control...and if you have better control you have better security" (cited in Atkinson, 2002).

RFID IN THE NEAR FUTURE AT THE AIRPORT

In a widely read article in *Scientific American*, Roy Want predicted that airline baggage tracking would be one of the first commercially viable RFID applications (Want, 2004). The potential market size is outstanding, as the world's airlines currently handle approximately two billion checked bags annually (Anonymous, 2003b). In the view of AIM Global (2004), with the proven accuracy and effectiveness of RFID-enabled baggage tracking, it may just be a

matter of time before the TSA mandates that such automatic identification technology-based systems be employed in the U.S. However, such mandates, whether in the U.S. alone or in conjunction with other civil aviation authorities worldwide, would raise a multitude of issues. These include who will bear the costs of such systems, the need for standards, and the need for international airlines that fly to the U.S. and/or interconnect with U.S.-based carriers to employ such RFID tagging.

Paul Coby, chief information officer of British Airways, believes that members of the airline industry need to work together to ensure that investments in technologies such as RFID will yield the fullest possible ROI and customer service benefits. He suggested that the International Air Transport Association (IATA) should play a leading role in driving this technology, so as to ensure that the industry adopts common information systems standards. Coby commented, "For technology to fully bring business change, the whole industry needs to move forward" (cited in Thomas, 2004). In June, 2004, Delta and United jointly proposed an RFID-specification for baggage to the IATA (Collins, 2004).

The need for a unique air transport standard is obvious for the not-so-distant future, looking to the day when luggage will contain items with their own RFID tags, say on Gillette razors, Benetton shirts, and items purchased from Target, Wal-Mart, Metro, or countless other retailers. Wal-Mart mandated the use of tags on merchandise it purchases from key suppliers by January 2005 (RFID Journal, 2003), thus prompting other retailers to follow suit, or at least begin investigating the technological investment such a move to RFID will require. For example, the retailer Boscov's (U.S.) sees customer service benefits in terms of reduced stock-outs, yet worries about the tag and infrastructure costs (Sullivan, 2004a). Tesco (UK)

announced plans to expand its RFID test project to include eight big-name packaged-goods manufacturers like Proctor & Gamble (Sullivan, 2004b). Others, like Federated's Lazarus store in Columbus, Ohio, and restaurants in Texas have seen improvements in customer service in terms of improved sales transactions (Coupe, 2003; Dunne & Lusch, 2005, p. 405). However, retailers will need to address the issue of consumer privacy, much like the airlines must do (Dunne & Lusch, 2005, p. 306; Lacy 2004).

It is even more important when one considers that the aircraft itself will likely have key parts tagged with RFID sensors in the near future. Boeing and Airbus are taking the lead in outfitting their new passenger jets, the 7E7 and A380 respectively, with RFID-tagged parts to provide a new level of historical and performance information on the key components. The two dominant commercial aircraft manufacturers are cooperatively working to produce industry standards, which is especially important since they share 70% of their supplier base (Tegtmeier, 2004). Likewise, Federal Express and Delta have pilot tested, equipping both flight deck electronics and engine parts, with RFID sensors (Brewin, 2004a). Thus, in only a matter of a few years, commercial airliners will be perhaps one of the most concentrated locations for RFID tags, making standards a necessity for avoiding problems with signal collision and information overload.

Tracking luggage with RFID may not be the only automatic identification technology we will see in use at the airport. By 2015, the International Civil Aviation Organization (ICAO) has proposed putting RFID chips in the over a billion passports worldwide. This move, while drawing fire from civil rights groups around the globe, may become a reality, all in an effort for the airlines and civil authorities to have better insights into who exactly to let on their aircraft (Jones, 2004). Likewise, the U.S. Transporta-

tion Security Administration has begun looking at how to use RFID-tagged boarding passes to improve airline security. The goals would be both to enhance airport security by giving facility security the ability to track passengers' movements within the airport and to speed passengers through airport security lines. The latter would be accomplished by linking the issuance of boarding passes to the proposed "registered traveler" program. This would allow frequent fliers who have been through a background check to be given specially tagged boarding passes, which they could then use to be directed through special "fast lanes" at security checkpoints (Brewin, 2004b).

The TSA is investigating the RFID-enabled boarding pass concept in concert with a number of other airport security initiatives in the United States. However, working in conjunction with the Federal Aviation Administration's Safe Skies for Africa Initiative, RFID-tagged boarding passes are already being deployed in an undisclosed number of African states (Brewin, 2004b). There are concerns however as to how this data will be utilized in airport security. From a practical standpoint, critics have scoffed at the jumble of data that would be created by trying to track thousands of passengers simultaneously in an airport. Privacy advocates also object to the invasiveness of the tracking, leading one to ask, "Are they going to track how long I spend in the ladies' room?" (cited in Brewin, 2004b).

There is also concern that airports, in their push to provide wireless access for patrons, may find that such Wi-Fi systems can conflict with RFID tracking innovations. In fact, in mid-2004, Northwest Airlines discovered that the wireless communication system used by its baggage handling operators was overwhelmed by a new Wi-Fi antenna installed by AT&T Wireless Services. The problem was alleviated after AT&T agreed to adjust its power levels,

but it seriously impinged on Northwest's own wireless systems for a time (Schatz, 2004).

A mid-September survey in 2004 by software supplier Wavelink found that approximately four out of five Frontline Conference and Expo attendees were currently piloting the technology or planned to do so in the next two years. Key concerns of the company executives included cost, lack of standards, and an early, untested market. Yet they expect adoption of the technology to grow, as it matures and benefits become reality (Gonsalves, 2004). However, as supply chain consultant Scott Elliff argues, all the new technology "simply isn't a substitute for superior business practices" (Elliff, 2004). The airline industry needs to remember this and, better yet, implement better business practices.

CONCLUSION AND FUTURE DIRECTIONS

This chapter has discussed the application of RFID in numerous airline applications across the world. The chapter has particularly discussed the advantages in using such applications for the benefit of all parties concerned. As has been shown, there is much promise for RFID to be applied in the airline industry to produce competitive advantage for airlines that are willing to implement the technology in a time of great competitive and economic turmoil in the industry. Both air carriers and airports themselves can leverage the technology to provide better customer service and heighten the security of air travel.

On a final note, moving to RFID may be the only way airlines may be able to even continue handling checked luggage in the future, both from a security and a cost standpoint. In fact, one company, the British-based low-fare carrier, Ryanair, has announced its intention to

eventually stop providing checked baggage service altogether. Michael O'Leary, Ryanair's maverick CEO, believes that by banning checked luggage, the cost of flying each passenger could be cut by at least 15%. This would be due to the elimination of the staff needed at check-in counters and in baggage handling operations. Not only would there be a direct cost savings for Ryanair, but there is the very real prospect for improved service, as passengers would get through the airport much faster and that their aircraft could be utilized more productively. The latter would be due to the quicker turn-arounds that the airline could achieve by not having to load outbound and unload inbound aircraft luggage holds (Noakes, 2004).

Over the next few years, Ryanair has planned to take steps to modify its passengers' mindsets regarding their baggage to encourage them to carry more of their baggage with them on board. The airline has already raised the weight limits for carry-on bags, while hiking its fees (up 17%) for overweight checked luggage. O'Leary even intended for the airline to begin giving passengers a small rebate if they choose to not check a bag sometime in 2005. While Ryanair's competition scoffs at O'Leary's luggage-ban plan, he notes that other innovations in the airline industry, including the elimination of paper tickets and Web-based travel booking, drew similar derision when they were first introduced (Noakes, 2004; Johnson & Michaels, 2004).

While the Ryanair gambit may prove to be prescient, for the near term, passenger luggage service will continue to be a cost of doing business for airlines. As such, Gene Alvarez, an analyst with Meta Group, predicts that RFID-based baggage tracking will become standard throughout the airline industry over the next decade (as cited in Brewin, 2004c). By assuring us that it is our black roller bag that ends-up on the luggage carousel at the end of our long journey home, airlines like Delta can seek a

competitive advantage through improved baggage service. In the end, we will likely see air carriers, airports, commercial aircraft manufacturers, and national transport and security agencies working cooperatively to smarten baggage handling through RFID tracking through common systems, while looking at other potential applications for automatic identification technology throughout the air transport industry.

REFERENCES

- AIM Global. (2004, January). Flying high. *RFID Connections*. Retrieved from <http://www.aimglobal.org/technologies/rfid/resources/articles/jan04/0401-bagtag.htm>
- Anonymous. (2003a). Tag tracking. *Airline Business*, 19(12), 14.
- Anonymous. (2003b, July). Luggage tracking trial by Delta Air Lines. *Smart Packaging Journal*, 11, 6.
- Anonymous. (2003c, November 10). Wal-Mart lays out RFID roadmap. *RFID Journal*. Retrieved from <http://www.rfidjournal.com/article/articleprint/647/-1/9/>
- Anonymous. (2004a, May 26). Hong Kong airport picks RFID baggage tracking. *Smart Travel News*. Retrieved from http://www.smarttravelnews.com/news/2004/05/hong_kong_airpo.html
- Anonymous. (2004b, May 20). Delta rolls out wireless baggage transfer system. *Smart Travel News*. Retrieved from http://www.smarttravelnews.com/news/2004/05/delta_rolls_out.html
- Atkinson, H. (2002). The allure of radio frequency. *Journal of Commerce Week*, 3(16), 28-30.

- Booth-Thomas, C. (2003). The see-it-all chip: Radio-frequency identification—With track-everything-anywhere capability, all the time—Is about to change your life. *Time*, 162(12), A8-A16.
- Bowen, B. D., & Headley, D. E. (2004). *Airline quality ratings 2004*. Retrieved from <http://www.unomaha.edu/~unoai/aqr/2004%20synopsis.htm>
- Brewin, B. (2003a, December 18). Delta has success in RFID baggage tag test, but a wide-scale rollout of RFID technology is being slowed by lack of money. *Computerworld*. Retrieved from <http://www.computerworld.com/mobiletopics/mobile/technology/story/0,10801,88390,00.html?f=x10>
- Brewin, B. (2003b). Delta's RFID trial run has airport predecessors. *Computerworld*, (June 23). Retrieved from <http://www.computerworld.com/printthis/2003/0,4814,82381,00.html>
- Brewin, B. (2004a). Delta to test RFID for parts tracking: Meanwhile, Boeing and Airbus are both pushing for common RFID standards. *Computerworld*. Retrieved from <http://www.computerworld.com/mobiletopics/mobile/technology/story/0,10801,93611,00.html?f=x10>
- Brewin, B. (2004b, April). TSA eyes RFID boarding passes to track airline passengers: Privacy groups view the idea as a "nightmare" for civil liberties. *Computerworld*. Retrieved from <http://www.computerweekly.com/articles/article.asp?liArticleID=127364&liFlavourID=1&sp=1>
- Brewin, B. (2004c, January 6). RFID bag-tag test proves a soaraway success. *Computer Weekly*. Retrieved from <http://www.computerweekly.com/articles/article.asp?liArticleID=127364&liFlavourID=1&sp=1>
- CNETAsia Staff. (2004, July 17). Japan firms to test radio-tagged luggage. *CNETNews.com*. Retrieved from http://news.com.com/2102-1009_3-1026860.html?tag=st.util.print
- Collins, J. (2004, July 2). Delta plans U.S.-wide RFID system: The airline carrier will spend up to \$25 million during the next two years to roll out an RFID baggage-handling system at every U.S. airport it serves. *RFID Journal*. Retrieved from <http://www.rfidjournal.com/article/articleview/1013/1/1>
- Collins, P. (2004). RFID: The next killer app? *Management Services*, 48(5), 20-23.
- Coupe, K. (2003, May 15). Customer loyalty initiative being tested by Texas restaurants. *MorningNewsBeat.com*. Retrieved from <http://morningnewsbeat.com/archives/2003/05/15.html>
- Deloitte. (2004). *Tag, trace, and transform: Launching your RFID program*. Retrieved from <http://www.deloitte.com/> (registration required).
- Dunne, P. M., & Lusch, R. F. (2005). *Retailing* (5th ed.). Mason, OH: South-Western.
- Elliff, S. A. (2004, September 6). RFID: Maybe not the "next big thing". *The Journal of Commerce*, 56(28), 46.
- Fein, A. (2004, September 8). Delta airlines to layoff thousands, outlines plan. *Axcessnews.com*. Retrieved from http://www.axcessnews.com/business_090804a.shtml
- Field, D. (2004). Radio waves. *Airline Business*, 20(7), 60-62.
- Gonsalves, A. (2004, September 23). Companies adopting RFID despite challenges, survey finds. *Information Week*. Retrieved from <http://www.rfidinsights.com/showArticle.jhtml?articleId=47902028&printableArticle=true>
- Hosenhall, M., & Kuchment, A. (2004, September 6). Crashes: Did "black widows" bring down the planes? *Newsweek*, 144(10), 6.

- Johnson, K., & Michaels, D. (2004, July 1). Big worry for no-frills Ryanair: Has it gone as low as it can? *The Wall Street Journal*, A1, A10.
- Jones, K. (2004). Are you chipped? *PC Magazine*, 23(15), 21.
- Jones, M. I., & Wyld, D. C. (in press). Smart tags + smart professors = smart students. *The Journal of the Association of Marketing Educators*.
- Kambil, A., & Brooks, J. D. (2002, September 1). *Auto-ID across the value chain: From dramatic potential to greater efficiency & profit—A white paper from the Auto-ID Center at the Massachusetts Institute of Technology*. Retrieved from <http://www.autoidcenter.org/publishedresearch/SLO-AUTOID-BC001.pdf>
- Kirby, J. (2003). Supply chain challenges: Building relationships—A conversation with Scott Beth, David N. Burt, William Copacino, Chris Gopal, Hau L. Lee, Robert Porter Lynch, and Sandra Morris. *Harvard Business Review*, 81(7), 64-74.
- Lacy, S. (2004, August 31). Inching toward the RFID revolution. *Business Week Online*. Retrieved from <http://www.businessweek.com/> (subscription required).
- McDougall, P. (2004, July 5). No more lost luggage. *InformationWeek*, 996, 14.
- McFarlane, D. (2002, May 1). *Auto-ID based control: An overview—A white paper from the Auto-ID Center at the Institute for Manufacturing of the University of Cambridge*. Retrieved from <http://www.autoidcenter.org/publishedresearch/CAM-AUTOID-WH-004.pdf>
- Morphy, E. (2004, July 2). Delta ups service bar with RFID luggage tracking. *CRM Daily*. Retrieved from http://www.newsfactor.com/story.xhtml?story_id=25711
- Murray, C. (2004, July 12). Airline clears RFID luggage tags for takeoff. *EE Times*, Retrieved from <http://www.embedded.com/showArticle.jhtml?articleID=22104613>
- Noakes, G. (2004, July 16). Ryanair bids to banish baggage. *Travel Trade Gazette*, 2623, 20.
- Perez, E. (2004, August 16). Flight upgrade: With Delta reeling, chief plans unusual bet on premium routes; as losses mount, Mr. Grinstein pairs cost cuts with extras; leather seats in coach; big risks in crowded skies. *The Wall Street Journal*, A1, A6.
- Reed Special Supplement. (2004). RFID: Powering the supply chain. *Logistics Management*, 43(8), R3-R16.
- Rothfeder, J. (2004, August 1). What's wrong with RFID? *CIO Insight*. Retrieved from http://www.cioinsight.com/print_article/0,1406,a=133044,00.asp
- Schatz, A. (2004, June 8). Airports clash with airlines over Wi-Fi. *The Wall Street Journal*, B1-B2.
- Sullivan, L. (2004a, September 20). Retail: General merchandising: Stores look for service with ROI. *Information Week*. Retrieved from <http://www.rfidinsights.com/showArticle.jhtml?articleId=47212239&printableArticle=true>
- Sullivan, L. (2004b, September 29). U.K. grocer expands RFID initiative. *Information Week*. Retrieved from <http://www.informationweek.com/story/showArticle.jhtml?articleID=48800020>
- Tegtmeier, L. A. (2004). RFID knowledge enabled logistics. *Overhaul & Maintenance*, 10(5), 24-28.
- Thomas, D. (2004, June 28). Airline industry needs to back IT. *Information World Review*.

The Next Big RFID Application

Retrieved from <http://www.iwr.co.uk/News/1156239>

Trebilcock, B. (2003). Ready to fly? *Modern Materials Handling*, 58(3), 40-42.

Tully, S. (2004, June 14). Airlines: Why the big boys won't come back. *Fortune*, 101-102, 104.

U.S. Department of Transportation, Office of Aviation Enforcement and Proceedings, Aviation Consumer Protection Division. (2004). *Air travel consumer report—August 2004*. Retrieved from <http://airconsumer.ost.dot.gov/reports/2004/0408atcr.doc>

Valentine, L. (2003, September 26). The new wireless supply chain. *CRM Daily*. Retrieved from http://cio-today.newsfactor.com/story.xhtml?story_id=22376

Want, R. (2004). RFID: A key to automating everything. *Scientific American*, 290(1), 56-66.

Weber, H. R. (2004, September 21). Delta pilots agree to recalls. *The [New Orleans] Times-Picayune*, C-1, C-7.

Woods, L. (2004). The fail-proof luggage finder. *Kiplinger's*, 58(10), 34-35.

Wyld, D. C., & Jones, M. I. (in press). A magic pill?: The emergence of radio frequency identification (RFID) technology in the pharmaceutical supply chain. *Journal of Pharmaceutical Marketing & Management*.

Wyld, D. C., Juban, R., & Jones, M. I. (2005). Dude, where's my cow?: The United States Automatic Identification Plan and the future of animal marketing. *Academy of Information and Management Sciences Journal*, 8(1), 107-120.

Chapter LIV

Identified Customer Requirements in Mobile Video Markets—A Pan-European Case

Torsten Brodt

University of St. Gallen, Switzerland

ABSTRACT

Due to a significant cost advantage, mobile multicasting technology bears the potential to achieve extensive diffusion of mobile rich media applications. As weak performance of previous mobile data services suggests, past developments have focused on technology and missed customer preferences. Mobile multicasting represents a radical innovation. Currently, little insight on consumer behaviour exists regarding such services. This chapter presents results of qualitative and quantitative field research conducted in three countries. It provides a continuous customer integration approach that applies established methods of market research to the creation of mobile services. Means-end chain analysis reveals consumers' cognitive reasoning and conjoint analysis drills down to the importance of service attributes. Desire for self confidence and social integration are identified key motivators for consumption of mobile media. Services should aim for technological perfection and deliver actual and entertaining content. Interestingly, consumers appreciate reduced but tailored contents and price appears not to be a superseding criterion.

INTRODUCTION

After its first years of existence, the still emerging mobile telecommunications industry is undergoing a period of fundamental change. Since previously high growth rates of voice revenues started to decrease, the industry is looking for additional sources of revenue, such as mobile

data services. However, the development of marketable services proves to be far more challenging than the one of stable, high-quality voice services.

Immature technologies are often blamed to be the reason for bad performances. Undoubtedly, the technological development is dynamic and, in fact, we argue that the intense focus on

technology push has been one key factor of the misfortune with mobile data services, as it detracts from customer needs. Furthermore, since vertical integration in the mobile telecommunication industry is low, product development is often organized in cooperative forms (Hagedoorn & Duysters, 2002). Coping with the complexity of innovation network management additionally detaches actors from actual customer needs.

Based on this, we see a need for a thorough understanding of the consumer behaviour side of mobile data services. Numerous studies have addressed issues of adoption and diffusion of mobile data services with the aim to identify diffusion barriers (e.g., Pedersen & Ling, 2003; Pousttchi & Schurig, 2004). However, such research seldom results in operational recommendations for companies on how to align their services with customer needs. We chose to focus on a specific range of services that exploit the investments in larger bandwidths and to develop a thorough understanding of the relations between service characteristics and fulfilment of customer needs and desires.

Since mobile multicasting services are based on a new technology and address a new market, they are termed a radical innovation (Veryzer, 1998). Thus, customer preferences can hardly be drawn from existing resources. By participating in the European “mobile multicasting service development and field trial project” MCAST (www.mcast.info), we were able to conduct the necessary market research.

Within a new product development process, customer integration is best realized after a first internal clarification of product ideas and possibilities, and subsequently after the technical engineering phase before market introduction (Gruner & Homburg, 2000). For this purpose, we integrated qualitative and quantitative methods to explore and formally describe customer needs. In the early stage we aimed to decrease uncertainty by conducting focus groups. We

complemented the results by conducting individual laddering interviews following the means-end chain framework (Gutman, 1982). With both methods we were able to obtain a complete set of service characteristics and the underlying cognitive reasoning. In the later stages of development, we conducted a prototype-based adaptive conjoint analysis to quantify relative importance and the preferred levels of service characteristics. These analyses were conducted in Switzerland, Israel, and Greece.

We claim three major contributions to extant research. First, our results provide information on what consumers expect of mobile video services and which reasons drive these expectations. Second, our results quantify the relative importance of service attributes, for example price vs. context dependency. Third, we provide a methodology on how customer needs for break-through mobile service innovations can be obtained. This enables a customer-centric development of radical innovations.

BACKGROUND—MOBILE MULTICASTING

MCAST’s multicasting technology enables cellular operators to use shared channel resources for broadcasting video and any other data over 2.5G and 3G networks. MCAST also yields a seamless roaming to WLAN networks. Therefore, MCAST aims at supporting cellular operators to establish affordable flat-fee services for end users and increase operators’ revenues per channel resource, allowing economic delivery of media to an unlimited number of cellular and WLAN devices.

Current Technology Constraints

Currently, rich media content can be delivered over cellular networks using unicasting (one-

to-one) technology. This has two major shortcomings: high delivery cost and limited cell capacity. Delivery cost is high, since each mobile terminal accesses a content server for on-demand content. When users view rich media content, their mobile terminals consume an excessive amount of bandwidth. This results in very high by-the-minute or by-the-packet charges. Due to limited cell capacity, unicasting of rich media can only support a limited number of subscribers at any given time. As the number of online users increases, additional bandwidth is required. Current technology performance, therefore, allows only poor service levels and implies lost revenues.

Challenge, Solution, and Opportunity

Multicasting technology is based on a one-to-many broadcast concept. It enables the delivery of identical content simultaneously to an unlimited number of subscribers. This allows services to scale to almost any number of users while having a manageable and limited impact on available bandwidth per cell. For the end user, multicasting represents a convenient way of accessing rich media content. In this sense, from a user as well as business model perspective, multicasting is believed to be a successful bearer for rich media content over 2.5G and 3G cellular networks.

Since there is currently no competing or ready-to-market technology that can provide multicasting services over 2.5G or 3G cellular networks, the MCAST research project moves on the forefront of technological development (Heitmann, Lenz, & Zimmermann, 2003; Northstream, 2002), and it will contribute to the ongoing standardization process of multicasting in the 3rdGeneration Partnership Project (3GPP). Alternative technologies like DVB-H required substantial investments in new net-

work infrastructure, and others like unicasting have an operational cost disadvantage. With its technological characteristics, multicasting is particularly suitable for rich media content (e.g., video, audio, gaming). Major market research institutions forecast the market potential of video services to nearly double that of audio services (e.g., Müller-Veerse, 2001) and a take up in 2005/06 (e.g., de Lussanet, 2003; Ovum Research, 2002). Based on this, our research focuses primarily on the delivery of video clips to mobile handsets.

EARLY-STAGE IDENTIFICATION OF CUSTOMER REQUIREMENTS

The early and qualitative part of customer integration employs focus groups to determine critical customer requirements as well as individual in-depth interviews following the means-end chain (MEC) methodology (Gutman, 1982) to understand the cognitive structures of decisions and the social motivation for requirements.

Focus Groups and In-Depth MEC Interviews: Background and Methodology

The focus group research was structured according to a theoretical concept for comprehensive and customer-driven product and service design: the OIL product design concept (Schmid, 2002). According to this, an evaluation of product expectations has to consider the levels of organizational design, interaction design, and logic design.

The organizational design level supplies the structural basis for the product design task. It answers the question of *who* and *what* is involved in the product use. Thus, in the case of a customer-oriented design of MCAST ser-

vices, user groups and content categories must be determined. The interaction design concentrates on the processes and interactions between the relevant elements defined in the organizational level. It thus answers the question of *how* the product will be integrated into everyday life. The logic design examines *why* users use a specific innovation. Based on this understanding of the decision process, the product's language and communication strategy can be designed.

The succeeding means-end chain (MEC) approach (Gutman, 1982) generates an understanding of customers cognitive structure. The MEC concept is partitioning this cognitive structure in three layer—that is, service attributes, needs, and values. In the market research and service design literature, the qualitative MEC analysis has been increasingly an object of scientific debate (e.g., Aschmoneit & Heitmann, 2002; Grunert & Grunert, 1995; Herrmann, 1996a, 1996b; Wansink, 2000). It is based on two assumptions: (1) values, defined as desirable end states of existence, are dominant in the formation of selection structures; and (2) people deal with the variety of services by forming classes to reduce decision complexity.

For the formation of classes, consumers consult perceived and anticipated consequences of their actions or decisions. They associate positive consequences, namely benefits, with certain decisions (Reynolds & Gutman, 1988). Personal values allocate a positive or negative valence to these consequences (Rokeach, 1973). Thus, a correlation between the concrete and abstract characteristics of a service, the functional and psychological consequences, as well as the instrumental and target values is assumed (Gutman, 1997). Since consumers form classes to simplify their decision-making process, relatively few values are connected to a larger number of consequences and attributes. In this hierarchy, the importance of values

determines the importance of consequences and attributes (Rosenberg, 1956).

Values represent beliefs about oneself and the reception of oneself by others. They are understood as universal, object-, and situation-independent convictions about desirable end states of life (Schwartz, 1994). The MEC framework is used to reveal the connections between time-stable values and product attributes directly relevant to decision making.

To obtain such results, the laddering technique with individual in-depth interviews is employed (Reynolds & Gutman, 1988). Research has shown that, on average, after 10 to 15 interviews, the number of additionally obtained consumer needs is decreasing radically (Griffin & Hauser, 1993). The technique reveals links between attributes, consequences, and values. The mentioned interactions between the obtained constructs were counted and entered into an implication matrix (not shown), a quantitative, tabular summary of the laddering interviews. This matrix provides the basis for the graphical representation in the form of a hierarchical value map (HVM), which displays the chains between values, benefits, and attributes, and their strengths (Herrmann, 1996a; Reynolds & Gutman, 1988).

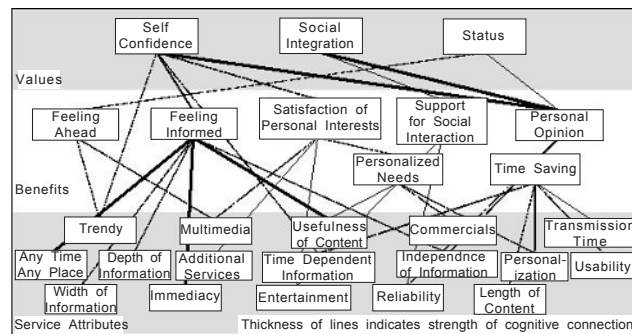
Focus Groups and In-Depth MEC Interviews: Results and Implications

In total seven focus groups were conducted, three in Switzerland and four in Israel. Participants were selected from two mobile operators' customer databases according to a screener questionnaire to find high-volume customers with strong interest in innovative services. Each group consisted of five to eight participants; discussions lasted 60 to 90 minutes. The identified issues relate to: (1) relevance and entertainment qualities of content;

Table 1. Exemplary focus group results—key requirements

	Group 1	Group 2
Content	• Good Editing of Content	• Availability/Quality of Content
	• Up-to-Date Content	• Up-to-Date Content
	• Local Content	• Width of Content
	• Fun	• Independence of Content
Technology	• Reliability of the Service	• Battery Consumption
	• Saving Functionality	• Screen Size
	• Picture Quality/Resolution	• Size of Device
	• Sound Quality	• Rapid and Secure Transmission
Service	• Personalization	• Anytime and Any Where
	• No Advertisement	• No Advertisement
	• Forwarding	• Forwarding
	• Easy to Operate	• International Roaming
	• Price	• Price
	• Customizability of Content	• Customizability of Content

Figure 1. Hierarchical value map (HVM)



(2) speed, visual quality, and reliability of technology; and (3) customizability of the service. We spare a detailed discussion of the focus group results and provide an exemplary overview of key requirements mentioned in two of the groups in Table 1.

For the MEC analysis, 30 innovators and early adopters were selected in Switzerland. The sample consisted of students and employees between the ages 20 to 40 of companies offering financial and consulting services. Interviews lasted between 30 minutes and one hour.

The obtained constructs complement on the one hand the requirements identified within the focus groups. On the other hand the MEC

approach allows structuring of the cognitive reasoning in an HVM (see Figure 1), summarizing service characteristics at the bottom, service benefits in the middle, and associated personal values at the top layer. For MCAST, two key paths of end user reasoning can be identified. One relates to information and self-confidence, and the other is associated with social integration:

- **Self-Confidence:** Being informed and deriving a personal opinion are among the main benefits associated with the reception of rich media content on a mobile device. That is, end users seek news content, enabling them to feel up to date at

any point in time. Three characteristics led to this benefit—the immediacy, the usefulness of the content, and the “any-time-and-any-place” characteristic. A service that follows this reasoning should not only provide updated information, but also ensure the contextual relevance of information.

- **Social Integration:** Consumers feel multicasting services may support them in achieving this goal by providing a basis for social interaction and the development of a personal opinion. While the latter greatly depends on the reliability of the service and the independence of the presented information, the support for social interaction also depends on the entertainment characteristics of the service.

The identified cognitive pathways provide guidance for service development. Immediacy, the relevance of content, and entertainment qualities should especially be taken into consideration. Winning companies will include the benefits of “Feeling Informed,” “Support for Social Interaction,” and “Forming Personal Opinions” to address the beliefs of consumers.

LATER STAGE CUSTOMER REQUIREMENT ANALYSIS

The preceding analyses show that customers consider a wide range of characteristics when evaluating a mobile multicasting service, which bears still too much complexity for service design. Therefore, we employed an adaptive conjoint analysis (ACA), a sophisticated customer research approach (Green, Krieger, & Wind, 2001; Hauser & Rao, 2002) to determine the weights of characteristics.

Adaptive Conjoint Analysis (ACA): Background and Methodology

The ACA allows identification of the relevance of service attributes and their levels—that is, it reveals the relative importance of different service attributes. The generated database allows the running of price sensitivity analysis for different product scenarios and an estimation of purchase probabilities (Johnson, 1991). Compared to other types of conjoint analyses, the ACA enables a dynamic adoption of a questionnaire according to given answers to preceding questions. This allows generation of robust results also for complex product offerings with a high number of attributes (Huber, Wittink, Johnson, & Miller, 1992; Orme, 1999) and ensures suitability for Web-based survey design (Dahan & Hauser, 2002).

Before implementing the ACA, the attributes under investigation have been reduced in an additional iteration step to 13 attributes. The objective of this step was not only to fulfil the conjoint requirements (of attribute independence, relevance, objective exclusiveness), but also to select attributes in conjunction with technological capabilities and business relevance. Accordingly, defined attribute levels are shown in Table 2.

The ACA was programmed using SSI Web of Sawtooth Software. The ACA questionnaire was hosted online and complemented by supplementary questions on general mobile usage behaviour and content requirements. This survey was conducted in three countries. In Switzerland 125 individuals have been invited from an academic database. Participants were required to be heavy users of mobile services. They were informed about the multicasting service by use of an animated prototype and in-depth information provided with an interactive CD-Rom. Participants were then asked to an-

Table 2. Attribute-level matrix

Attribute	Level 1	Level 2	Level 3
Length of Content	• Max. 30 sec.	• Max. 1 min.	• Max. 2 min.
Number of Clips per Day	• 5	• 10	• 15
Premium Content	• Available	• Not Available	•
Subscription Fee (•)	• 3	• 6	• 9
Forwarding	• Via MMS	• Not Possible	•
Ensured Transmission	• Retransmission	• Clips Lost	•
Supplemental Internet Service	• All Clips Online	• Missed Clips Online	• No Clips Online
Advertisements	• Yes	• No	•
Notification on Missed Clips	• No	• Per SMS	• Per MMS
Number of Content Categories	• 5	• 10	• 15
Number of Clips in MCAST Inbox	• 3	• 5	• 7
International Roaming	• Available	• Not Available	•
Location-Based Content	• Available	• Not Available	•

swer the online survey. After data was cleaned to ensure data robustness, 103 data sets were used for analysis. In Greece and Israel the participants have been recruited from the project partners' databases. The main difference was that users in these countries had the opportunity to use the service in the life network for a duration of four weeks. After data cleaning, the analysis contained 67 participants in Greece and 97 in Israel.

Adaptive Conjoint (ACA): Results and Implications

Questions complementary to the actual ACA asked for the general background of participat-

ing individuals (e.g., demographics, telecommunication behaviour). Among others, these questions confirmed the interest in news and location-specific content. Furthermore, video content proves to be the most preferred format.

The following section selectively documents the quantitative conjoint results. Data reveals that in all cases, five service attributes influence almost 50% of the consumer decision (see Figure 2). Not surprisingly, price concerns yield a high score of 13.8% in Israel, 10.6% in Greece, and 11.7% in Switzerland. However, attribute scores are rather evenly distributed, and given an adequate price span and a flat fee pricing model, price appears to be not an overriding decision criterion for consumers.

Figure 2. Cross-country comparison of attribute importance from ACA

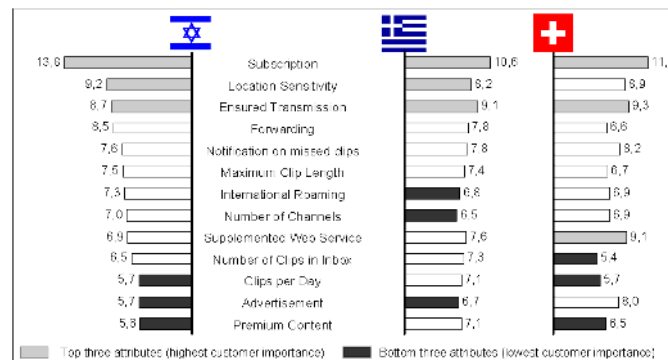


Figure 3. Attribute levels for clip length and number of clips per day (Switzerland)

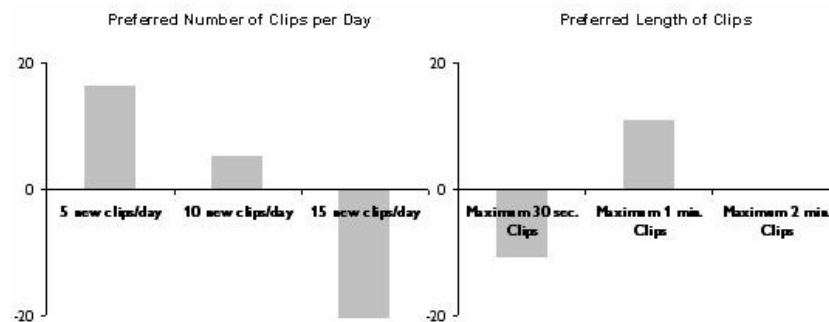


Figure 2 depicts the attribute importance in consumer decision making for the three countries. Certain similarities can be identified—for example, the high weight of the item “Ensured Transmission,” which points to the importance of a technically flawless service. Since multicasting services are broadcast to a group of subscribers once and simultaneously, it might happen that a few subscribers do not receive the content due to handset unavailability or interrupted transmission. Users are concerned to lose out on these clips and therefore strongly require the notification and the back-up through supplemental Internet services.

Taking a closer look at single attributes (here we chose the data for Switzerland) and their levels reveals interesting aspects of the willingness to consume mobile data services. As documented in Figure 3, subscribers rather prefer a reduced number of (five) clips per day combined with a maximum clip-length of one minute. This behaviour relates to the concern about content relevance, but also about technical capacity (e.g., transmission speed and memory capacity) mentioned during the preceding qualitative surveys.

Analysing the findings of the three country-specific surveys on an aggregate level reveals four main patterns of consumer behaviour regarding mobile video services. As shown in Figure 4, these patterns relate to (1) the proven

existence of a willingness to pay, if the price is controllable, preferably a flat fee. (2) The second pattern describes the users’ preference for a reduced but tailored mobile video offer; that is, despite flat pricing, users do not always opt for maximum of outputs. This behaviour is rooted in concerns for relevance, information overload, and technical capacity, as also shown in the MEC-analysis. (3) The very advantage of mobile technology of delivering services “anywhere-anytime” is also a valuable selling point for mobile video services. That is, companies must develop intelligent means to satisfy the need for current and contextualised (personalised/localised) services, without destroying the scale effects of mass-broadcasting. (4) Precision is precious—this pattern represents the users’ concerns about technical reliability rooted in past cognitive dissonance and disappointing experiences, and it implies a call for command of technology.

CONCLUSION AND BUSINESS BENEFITS

By reporting insights in terms of methodology and identified customer preferences regarding mobile-rich media services, we address the lack of customer knowledge in marketing practice and research in the mobile media industry.

Figure 4. Evidence for behavioural patterns in observed countries

Insight	Rationale			
Price flat "but" stable	- Given adequate price span and flat fee pricing, analyses reveal price achieves important but not dominant weighting - Advertisement not perceived as disturbing and personalization of ads can increase revenue potential, unique to the mobile business model			
Reduce to the max	- Despite flat pricing, users don't always opt for max. outputs - Research shows subscribers prefer reduced but tailored offer - Rooted in concerns for relevance, info-overload and technical capacity - Applies to clips per day, clip length, storage on handset			
Anywhere anytime and interactivity	- Anywhere-anytime advantage exists and must be leveraged: Companies must satisfy need for current & contextualised (personalised / localised) info services to benefit from mobile channel - Challenging: Trading off the cost advantage of (mass-) broadcasting			
Precision is precious	- Items like "ensured transmission", and "suppl. web services" impact more than 25% on consumers' evaluation - Mirrors technical reliability concerns and implies a call for command of technology - rooted in past cognitive dissonance and disappointing experiences			

Strength of empirical evidence
 Strong  Weak

While dealing with the development of a leading-edge multicasting technology, we deployed a set of sophisticated tools for customer integration along the development process. For customer research science, we show a methodology, on how customer needs for breakthrough mobile service innovations can be obtained in a way that generates results, which can be easily communicated within single companies and across innovation networks. With the growing importance of cooperative product development, investigations on the latter, such as a joint customer integration and its qualities, will be an area for future research.

For management, our quantitative empirical results imply precise insights for superior mobile multicasting service design. Additionally, the identified cognitive reasoning of consumers provides input for general communication and marketing strategies. We show that most importantly, management needs to master the doubts on technology performance, and that mobile content must be tailored. The latter point complicates the marketing challenge as it trades off the multicasting cost advantage. For marketing and communication strategy, we have identified that the consumers' desire for self-

confidence and social interaction should be addressed.

NOTE

An earlier version of this chapter appeared in: Cunningham, P., & Cunningham, M. (Eds.). (2004). *E-adoption and the knowledge economy: Issues applications, case studies* (Vol. 1, pp. 50-58). Amsterdam: IOS Press.

REFERENCES

- Aschmoneit, P., & Heitmann, M. (2002). Customer-centred community application design. *The International Journal on Media Management*, 4(1), 13-21.
- Dahan, E., & Hauser, J. R. (2002). Product development—Managing a dispersed process. In R. Wensley (Ed.), *Handbook of marketing* (pp. 179-222). Thousand Oaks, CA: Sage Publications.
- de Lussanet, M. (2003). *Limits to growth for new mobile services*. Cambridge, MA: Forrester Research.

- Green, P. E., Krieger, A. M., & Wind, Y. (2001). Thirty years of conjoint analysis: Reflections and prospects. *Interfaces*, 31(3), 56-73.
- Griffin, A., & Hauser, J.R. (1993). The voice of the customer. *Marketing Science*, 12(1), 1-17.
- Gruner, K. E., & Homburg, C. (2000). Does customer interaction enhance new product success? *Journal of Business Research*, 49(1), 1-14.
- Grunert, K. G., & Grunert, S. C. (1995). Measuring subjective meaning structures by the laddering method: Theoretical considerations and methodological problems. *International Journal of Research in Marketing*, 12(3), 209-225.
- Gutman, J. (1982). A means-end chain model based on consumer categorization processes. *Journal of Marketing*, 14(6), 545-560.
- Gutman, J. (1997). Means-end chains as goal hierarchies. *Psychology and Marketing*, 14(6), 545-560.
- Hagedoorn, J., & Duysters, G. (2002). External sources of innovative capabilities: The preferences for strategic alliances or mergers and acquisitions. *Journal of Management Studies*, 39(2), 167-188.
- Hauser, J. R., & Rao, V.R. (2002). *Conjoint analysis, related modeling, and applications*. Unpublished manuscript, MIT Sloan, USA.
- Heitmann, M., Lenz, M., & Zimmermann, H.-D. (2003). *Preliminary user needs analysis for MCAST*. St. Gallen, Switzerland: MCM Institute.
- Herrmann, A. (1996a). *Nachfrageorientierte produktgestaltung: Ein ansatz auf basis der "means end"—theorie*. Wiesbaden, Germany.
- Herrmann, A. (1996b). Wertorientierte produkt—und werbegestaltung. *Marketing ZFP*, 18(3), 153-163.
- Huber, J., Wittink, D. R., Johnson, R. M., & Miller, R. (1992). Learning effects in preference tasks: Choice-based versus standard conjoint. In *Proceedings of the Sawtooth Software Conference*, Ketchum, ID (pp. 232-244).
- Johnson, R. (1991). Comment on adaptive conjoint analysis: Some caveats and suggestions. *Journal of Marketing Research*, 28, 223-225.
- Müller-Veerse, F. (2001). *UMTS report—an investment perspective [online]*. Retrieved August, 2002, from <http://www.durlacher.com>
- Northstream. (2002). *The competitive landscape of mobile video on demand [online]*. Retrieved February, 2002, from <http://www.northstream.se/21/>
- Orme, B. (1999). *ACA, CBC, or both?: Effective strategies for conjoint research: Sawtooth software*. Sequim, WA.
- Ovum Research. (2002). *Ovum forecast: Global wireless markets*. London.
- Pedersen, E., & Ling, R. (2003). Modifying adoption research for Mobile Internet service adoption: Cross-disciplinary interactions. In *Proceedings of the 36th Hawaii International Conference on System Sciences 2003*, Hawaii (pp. 534-544).
- Pousttchi, K., & Schurig, M. (2004). Assessment of today's mobile banking applications from the view of consumer requirements. In *Proceedings of the 37th Hawaii International Conference on System Sciences 2004*, Hawaii (pp. 184-191).
- Reynolds, T. J., & Gutman, J. (1988). Laddering theory, method, analysis, and interpretation. *Journal of Advertising Research*, 28(1), 11-31.
- Rokeach, M. J. (1973). *The nature of human values*. New York: The Free Press.

Rosenberg, M. J. (1956). Cognitive structure and attitudinal affect. *Journal of Abnormal and Social Psychology*, 22, 368-372.

Schmid, B. (2002). *Kommunikations—und medienmanagement*. Unpublished manuscript, St. Gallen, Switzerland.

Schwartz, S. H. (1994). Are there universal aspects in the structure and content of human values? *Journal of Social Issues*, 50(4), 19-45.

Veryzer, R. W. (1998). Discontinuous innovation and the new product development process. *Journal of Product Innovation Management*, 15, 304-321.

Wansink, B. (2000). New techniques to generate key marketing insights. *Marketing Research*, 12(2), 28-36.

Chapter LV

Applying Mobility in the Workforce

Bradley Johnstone
BK Solutions, Australia

Khimji Vaghjiani
BK Solutions, Australia

ABSTRACT

There have been significant advances in mobile technologies in recent years. The euphoric technology void left by the dot-com crash in early 2000 soured many technology users; however mobile computing has provided much needed enthusiasm for both technologists and business users. In this chapter we focus on aspects of mobile technology, from both a business user perspective and a technology view point. Aspects such as total cost of ownership, return on investment and capital investment have been discussed from a financial perspective. Technical aspects of running and maintaining a mobile technology infrastructure have also been explored. The chapter concludes with a review of potential areas of application for mobile technology. The area discussed is mobile technologies in banking; however, many of the aspects covered could easily be applied to any other business vertical. Finally, this chapter is not meant to be a holy grail for mobile computing. It is simply a glimpse of the need to explore the power of this emerging technology.

THE MOBILE WORKER

This chapter discusses the application of mobile technologies in the business processes related to a mobile worker, with specific emphasis on the financial and banking industry. The term mobile worker has been used for many years to describe a worker who travels to various locations in order to conduct their business. This

could describe a mechanic going to the car that needs repair, a courier delivering parcels, or even a sales representative who travels in order to showcase and sell his or her products. While these workers have been around for a long time, the term mobile worker has increasingly come to represent a mobile workforce well equipped with mobile technologies and associated devices.

The needs of the aforementioned workforce, coupled with recent advances in telecommunications technologies and infrastructure across the globe, has resulted in abundant opportunities for many businesses to create new mobile-enabled business processes that were hitherto unimaginable. For example, it is now possible to transfer an enormous amount of data at speeds that allow devices and the applications running on these devices to interact in real time with each other irrespective of time and location. The Internet facilitates connection of businesses across geographical and time barriers resulting in increased collaboration demanding innovative and technologically savvy business processes (Unhelkar, 2003). For example, many small business owners are totally eschewing physical offices and, instead, operate their businesses almost entirely on the road by fully utilizing the capacities of their mobile telephones and personal digital assistants (PDAs). Medium to large businesses also have a need to customize those business processes that can incorporate mobile technologies. However, the manner in which businesses and people within businesses utilize mobility depends on the type of industry in which they operate and the specific organization's own methods of doing business. The deployment of a mobile workforce has unique challenges that also depend, to a large extent, on the individuals that comprise the workforce. UK Company Softlab Ltd. (2004) suggests that mobile workers fit into one of three core groups requiring a mobile solution:

- management,
- sales and marketing personnel, or
- customer service representatives.

Managers are able to make key decisions and send and receive important correspondence via laptops and PDAs anywhere within mobile network coverage. Sales and marketing

personnel can close deals with their clients and process the order while still at their client's premises. This alone ensures that the level of customer service is higher by allowing the worker to spend more time in front of the customer and not having to wait until they are back at the office to fulfil the order.

Customer service representatives can be armed with many different types of mobile devices designed to enhance the customer experience. An example in the banking sector is the ability to approach a customer waiting in line at a branch and quickly processing their request by connecting wirelessly to the bank's internal network via a tablet PC (Kuykendall, 2004).

The mobile workforce will grow rapidly as the next wave of mobile solutions arrives. We can expect to see a much wider deployment of applications to a greater number of workers from different industries. The applications that mobile workers use will vary depending on the business need. The 2003 findings of UK Company QNB Intelligence, who interviewed IT directors across Europe to gauge what applications they would use in their business if they were to deploy a mobile workforce, showed:

- 88% required general office applications
- 56% required applications to automate the sales processes
- 46% required tools to assist field agents
- 36% required applications to assist delivery and collection of goods
- 24% required management recording tools

Some sample applications within the above categories are explained further.

General office applications include tools such as e-mail and calendar events. A worker can have access to his or her e-mail or schedule via a laptop or other handheld device. This enables him or her to have the latest information on

Applying Mobility in the Workforce

hand at all times, including receiving changes to schedules. This solution requires little investment, as the devices and software are readily available.

Products that automate the sales process along with customer relationship management (CRM) tools are being deployed in enterprises all over the world in order to effectively track customer habits and provide better customer service. It seems only natural that these tools should also extend to the mobile worker. In fact many CRM providers have already built mobile workforce functionality into their systems. The salesperson who has access to all of the clients' data is in a better position to answer any queries a customer may have and also allows for longer face-to-face time with a client.

Tools to assist field agents can also ensure that agents have their schedules arranged and updated throughout the day by the head office. They also ensure agents' access to a company's inventory, allowing them to field customer orders on the spot with up-to-the-minute inventory levels. The agent may use a tablet PC that displays an order form so by the time they are ready to leave the customer, the order is already being processed at the warehouse, allowing for faster delivery.

Delivery and collection of goods improve greatly with the inclusion of mobile technology (Egan, 2003). While delivery may be predetermined based on the parcels a driver may have on board, calls for collections come in at different times. New collection information can be pushed to the driver and placed within the list in the most efficient order. With the use of GPS systems, a mobile device can even give the driver directions to the next port of call.

Management recording tools can be particularly useful for consultants who charge based on effort—they are able to record start and finish times on their handheld devices, which are then relayed back to their office's

billing system for automatic processing (Bailey, Buist, & Vile, 2003).

Of course there are so many more applications available, and the above represents only a mere fraction of the mobile solutions that are already available with these five main categories.

FORMS OF MOBILE WORKER

As previously discussed, a mobile worker is usually considered to be a person who conducts their business by travelling to varied locations. It is important to note that an employee may in fact remain in one fixed location and still be considered a mobile worker. The definition can therefore be extended to include workers who can connect to a company's system with a device that can be easily moved from one place to another. A typical employee will work at a single location with a device that is attached to the company's local area network via a physical cable; this of course limits their ability to perform tasks on the go. Others may require the ability to connect to a network or Internet from any location outside the bounds of physical cables.

In assessing what sort of technology should be deployed, we must consider the employees' roaming habits. It is possible to group mobile technology into four basic categories (Gessel, 2001):

- Stationary at one location
- Stationary at many locations
- Many locations and moving
- Concealed and moving

Stationary at One Location

People who fall under this heading generally work from one location either at or near their desk. These workers may use a laptop which is

easily moved or a handheld device such as a PDA or a tablet that accesses the company's wireless network.

Stationary at Many Locations

Employees who find they work in multiple locations may fall under this category. Similarly to the above scenario, these users are in the range of the company's LAN or wireless network generally by being at a premises operated by the business. An example in a banking context would be a financial advisor who operates at different company locations on different days and requires a connection to the network from a device they carry around to all of the locations.

Many Locations and Moving

Many workers find that they perform tasks in many different locations and are subject to many working environment constraints like physical access to a company's network. They may opt to use handheld devices or mobile telephones which allow them to remotely connect to the company's network via a mobile telephone carrier (Broersma, 2004). In a banking context, we might consider a financial advisor visiting a client at their home with the ability to look at the client's online file when and as needed.

Concealed and Moving

There are mobile application scenarios that require no human interaction. A primary example is a warehouse system that tracks stock through using radio frequency identification (RFID) tags. As human involvement is not required, the tasks are generally high speed and non-complex. A typical warehouse RFID system tracks company inventory levels upon request or at pre-determined intervals.

CONSIDERATIONS BEFORE GOING MOBILE

For many years, U.S. courier company FedEx has been using custom-made equipment that allows drivers to record pickups and deliveries on a handheld device. This data is stored until the driver arrives back at base where it can be uploaded to the company's network (Egan, 2003). This legacy technology is being replaced with handheld terminals modeled on the Pocket PC operating system that incorporates General Packet Radio Services (GPRS) that can report changes as they happen. This allows FedEx to accurately track their drivers in order to give better estimated times of arrival to clients.

This mobile solution is returning clear benefits to both the company and their customers, some of which include (Urich, 2002):

- a reduction in paperwork,
- shortening business process cycle times,
- reducing labour requirements, and
- freeing up time to allow for greater productivity.

The benefits of a more efficient and better organized workforce returns real cost savings as well as competitive advantage for the organization.

A business case should justify the investment in new mobile technology. It must be clear that there will be an improvement in the businesses current processes gained from adopting the new technology. Technological solutions have in the past failed because procuring departments thought the technology was "sexy" and they failed to purchase based on a benefits or needs basis. No investment should be made without a strong business case justifying the investment. Preparing a business case for mobile strategies has so far withstood intense scrutiny by company finance departments who demand clear justification on spending (Gold,

2005). A 2003 report by Mobile Competency Founder Bob Egan suggests that when preparing a mobile strategy, management should assess the total financial impact including both *Total Cost of Ownership* and traditional *Return on Investment* analyses.

Total Cost of Ownership (TCO)

When assessing the TCO of a mobile solution, it is important to consider and ensure optimal investment in each of the following four areas:

- Capital Expenditure
- Operations
- Support Services
- Administrative Operations

Capital Expenditure

In deciding where to invest, management should try to standardize the solution wherever possible. This will allow for greater bulk discounts on equipment and services which will in turn lower the ongoing support costs. Deployment costs are also likely to be lower when a standardized solution is implemented.

Operations

How will the implementation of a mobile solution affect the productivity of an end user? Any increase in productivity could potentially reap a large company enormous savings when spread across its entire workforce. The increase in productivity may also equate more face-to-face hours with a sales force and their customers.

Egan (2003) discusses the case of a global bottling group which enabled their sales staff with handheld devices that included a smart selling application. As a result of this new

technology, the sales staff close rates dramatically improved. The system provided sales staff with prompts that identified up-selling and cross-selling opportunities, and even helped determine future demand.

Support Services

The success and acceptance of a mobile program by employees may hinge on the available technical support that is provided (Egan, 2003). Standardizing the system should reduce any negative impact by allowing IT support to become familiar with all offerings which will allow them to quickly resolve any issues that staff may have. Through adopting a standardized approach early, change management processes will be easier to control.

Administrative Operations

Administration costs may also improve over time as processes become more efficient. Traditional mobile workers spend their day filling in paperwork only to have a team verify and process the results at a later time. If they used a mobile device, the administration team could have access to their data in real time or some processes could become fully automated.

Return on Investment (ROI)

There are many ways to calculate ROI for new solutions. Some outcomes are predictable and others harder to ascertain without specific use cases. When calculating ROI for a mobile solution, the two benefits below may assist the calculation:

- Efficiency Gains
- Improvements in Staff Productivity and Effectiveness

Efficiency Gains

Speeding up a specific task or application can be a simple way of generating a return. For example, providing financial advisors with the ability to view client data while on site will allow them to keep the customer focused, which may not be the case had the advisor needed to make a phone call back to base which of course will involve another bank employee. This process is likely to keep them in control of the session instead of breaking the momentum of the meeting to retrieve the client's data. The speeding up of this process will allow the advisor to present all facts and better answer questions, which will put them in a better position to get the client's business.

Improvements in Staff Productivity and Effectiveness

Field agents can improve their effectiveness by having access to e-mail, calendar events, and contact data while on the road. Using a laptop connected to a mobile phone for this task is not a new concept, but it is cumbersome. Accessing this information on a handheld device connecting via a mobile telephone network is likely to receive more use. It is also likely to be a more economical solution as the GPRS network is charged on data transferred, whereas the laptop and mobile phone solution will incur charges based on the amount of time the user is connected (Deshpande & Gilbert, 2002).

MANAGEMENT BARRIERS TO MOBILITY

Other than the typical budgetary restraints, there are a number of other barriers whose existence can generally be attributed to a manager's perception of the effectiveness of a mobile workforce or other trust issues.

Many see benefits in adopting mobile technology that assists workers who are on the road, or improves the efficiencies of staff at a company location. There is another group that needs consideration, however: the flexible worker. This is a worker who is granted the opportunity to work from a location other than the office for mainly work life balance reasons or to make savings in office rental costs (Sweeney Research, 2002). The flexible worker tactic may also be used to retain quality staff, and affording them the luxury of flexibility allows them to better handle working pressures.

Some of the key areas that can inhibit a mobile worker are:

- Managers' Lack of Trust in Mobile Workers
- Hostility Between Mobile and Non-Mobile Workers
- Motivation Issues

Most of the above issues could best be addressed by managers setting performance evaluation techniques that would allow them to monitor the workers' output in terms of quality and productivity.

Managers' Lack of Trust in Mobile Workers

A paper by Sweeney Research (2002) revealed that more than 50% of managers surveyed were less likely to trust workers who had a flexible arrangement and could elect to work from home. This attitude is perhaps the greatest inhibitor to realising a widespread flexible workforce.

Hostility between Mobile and Non-Mobile Workers

Another issue which may arise is the mistrust from a non-flexible worker towards the flexible

worker. The traditional workers surveyed by Sweeney suggested that they felt that flexible workers did not work as hard as they did.

A common issue was that unsupervised flexible workers were perceived to be more likely to conduct their personal business during working hours and therefore work less productively.

Motivation Issues

Managers surveyed by Sweeney felt that as they did not have traditional control of flexible workers, they would find motivating them to be a very difficult task. A manager could not control the distractions that a flexible worker faced such as watching television or conducting their personal business when the employee should be working.

Before adopting a mobile solution for the workforce, it is crucial that the above areas are considered and a clear strategy is in place to stop these areas from becoming an issue within the organisation. Failure to address this early will lower morale and breed mistrust between workers. This will have a direct impact on quality of customer care, and the bottom line will suffer. With clear ground rules set in stone early, going mobile can produce enormous gains in productivity, and improvements to the workers' quality of life as well as customer satisfaction rates.

SECURITY CONSIDERATIONS

Security is important in all applications, but becomes an essential consideration when applications are used in the field. A mobile product should support a wide variety of authentication and encryption standards to integrate easily into an existing infrastructure. This area should be addressed by managers before deploying a mobile worker, and risks must be

identified and mitigated before any investment is considered.

Wireless Policy

Companies are very protective of the data on their networks. Deploying a mobile workforce brings a range of new security challenges that must be addressed (Gallagher, 2004). The solution starts with a top-down policy that addresses why wireless is being used, what the business objectives are, and what policy governs the entire company in this area.

The wireless policy must demand that workers strictly adhere to the standards contained within and also govern how workers will protect the devices within their possession.

Details of what to do to prevent loss of a device or immediate action required in the event that the device is lost should be well known to all users. If the device stores sensitive information, then a high level of security is required. A high level of security should involve having all files on the device password protected and preferably encrypted. There are many different methods that can be used to protect a company's data including encryption, data wipe technology, secure transfer, firewalls, and a worthy protectionist login regime.

Encryption and Secure Data Transfer

The mobile workforce should worry IT security specialists who need to protect company data. Mobile solutions should incorporate security standards that encrypt data between the server and mobile device. Session-based keys should be used with any encryption method, and a secure public key private key exchange should be employed to ensure communications integrity (Hildebrand, 2004).

Processing transactions securely on the Web means that we need to be able to transmit

information between the client and server in a manner that makes it difficult for other people to intercept and read. An example of a secure data exchange is through the use of Secure Sockets Layer (SSL) sessions.

SSL sessions work through a combination of programs and encryption decryption routines that exist on the hosting computer and in client browser programs such as Microsoft's Internet Explorer. Below is a high-level description from Intel that explains how an SSL session is negotiated between server and client:

1. Browser checks the certificate to make sure that the site you are connecting to is the real site and not someone intercepting.
2. Determine encryption types that the browser and Web site server can both use to understand each other.
3. Browser and server send each other unique codes to use when scrambling (or encrypting) the information that will be sent.

The browser and server start talking using the encryption, the Web browser shows the encrypting icon, and Web pages are processed secured.

In addition to SSL sessions, it is important to understand that wireless devices can be more of a target for theft or are easily lost. Companies concerned that the device may go missing or fall into the wrong hands should consider the sensitivity of the data on the device and assess that risk. Organizations should have a plan in place to guard against this scenario. Wireless devices can encrypt all data written on its hard drive, rendering it useless to anyone who tries to read the data, whether from the device itself or by any other means (such as mirroring the hard drive).

Encryption Toggling

Depending on the circumstances, encrypting data flow may or may not be critical. When connected to a LAN inside the office, protected by the company firewall, encryption may not be as essential. In addition certain user groups may deal with less sensitive information that does not require encryption thus speeding their communications sessions.

Allowing the system to automatically or manually turn the encryption feature on or off for some applications will ensure that the network is secure and minimizes performance impact according to Intel's 2004 article "Wireless Security Best Practice." This process is known as *encryption toggling* and is designed to enhance the user experience in less secure environments. It is important however to fully assess the risk and sensitivity of data and user permissions to change the settings affecting toggling.

Multiple Authentication Modes

Any new solution should support all current and future company authentication modes. End users should be able to authenticate to the server using existing user IDs and passwords, whether connecting via LAN, WAN, or GPRS.

In addition, adding a new user or a new device by an administrator should require no additional IT intervention.

Sniffer Solutions

Companies are not alone in their fight to secure data over wireless networks (Intel, 2004). Sniffer technologies are available for enterprises that need to lockdown systems on mobile devices. Many of the benefits of wireless technologies also create risk. While the flexibility of being able to operate in any location is the primary attraction of wireless networks, the

need for security must also be considered. It is important to have appropriate security to ensure that data is stored properly, travels properly, and is protected from people who should not have access to it, but is still accessible to those who do need it.

Sniffer products function in a wireless environment like LANs or WANs. A wireless sniffer card in a laptop can detect and verify that your encryption protocols are being used and can find rogue access points, enabling you to shut them down if necessary.

MOBILE SOLUTIONS FOR THE FINANCIAL INDUSTRY

Financial service providers have invested heavily in applications with a consumer focus. While investment in this sector will continue, banks are looking at improving internal efficiencies and providing new tools for employees to assist in delivering an exceptional customer service experience.

Financial institutions are in a broadly defined category, and there are many different employees that can benefit from mobile technologies. Some of those who would benefit are financial advisors, fund wholesalers, insurance agents, management, brokers, risk inspectors, investment bankers, analysts, customer service representatives, loan officers, and many other disciplines (Intellisync, 2003). Some of the ways that mobile technology could be used by a bank are detailed as follows.

Branch Applications

The next step in wireless banking is likely to focus on technology used within the bank, including wireless Internet access and access to company information. These types of wireless capabilities can increase employee pro-

ductivity and customer satisfaction. A roving teller can use a number of different devices, like a portable notepad or laptop or even a PDA, and can greet the people as they come in the door. For example, if customers want to find out if a check has cleared, they can find out without needing to wait in line, thereby reducing waiting time for other customers.

Implementing a wireless network within the bank can also prove to be cost effective. Banks looking to expand can benefit from installing a wire-free network. If a bank is going to open up new branches, it will likely be more cost effective to deploy a wireless network vs. a wired network (Ramsaran, 2004). However, banks may not be so quick to move to a wireless network because of security issues and standards protecting customer data. Manufacturers claim that this is because they are not fully informed of the steps taken to secure an internal wireless network. They also claim that there is no more reason to be concerned about that than any other wired network security because it has the same protection.

E-Mail

An obvious first step in extending mobile communications to the mobile workforce is empowering them with handheld devices that can send and receive e-mail. Response times to internal and external customers are greatly improved, as workers can address e-mail as soon as it arrives just like in a fixed location.

The type of device used will depend on the requirements of the worker. If simple text mail messages are being sent, then any device from a mobile phone to a PDA would be fine. In the case of large attachments, anything from a PDA up to a laptop connected via a mobile phone would be required. In any event the type of device deployed must match the business need and be able to perform adequately prior to any investment.

Executive Dashboards

With the financial services moving at a frenetic pace, executives may struggle to keep abreast of what is happening within their organisation. This information can easily be provided on a handheld device. A dashboard shows key performance metrics that will assist the manager in making informed decisions with up-to-date information.

Software Maintenance

IT teams will be able to dial in to the company's mobile device and perform software upgrades remotely. These measures will save time and money since the employee will not have to travel to hand in the device. The list of benefits continue, with more expensive alternatives avoided such as having to burn and send CDs or sending out a field technician to perform the install or update.

Sales Force Automation (SFA)

Account managers with extended SFA applications should boost productivity by speeding up each process along the sales cycle. Allowing anytime access to vital corporate data allows for improved customer responsiveness, better tracking of customer activity, increased customer acquisition, and enhanced communication with the corporate office.

Competitive and Product Information Delivery

Remote and mobile workers benefit greatly from being able to receive frequent updates to competitive analysis reports and regulatory requirements during remote selling opportunities. It is critical for all remote personnel to have the latest product and policy information to avoid making false promises to customers.

Wholesale and Retail CRM Applications

The most important goal for wholesale and retail sales representatives is to cultivate meaningful relationships with clients. To succeed, representatives need easy access to large amounts of contact and sales data, enabling them to make more sales calls and improve the quality of the calls. This data can also assist with up-selling and cross-selling opportunities.

Quoting for Insurance Agents

The abundance of paper-based processes gives the opportunity for a high percentage of submission errors. Automating the task of filling out insurance application forms, and allowing agents to provide easy quoting and sales tracking will improve an organization's close rate.

Commission Reporting and Tracking

The complexity of tracking compensation and incentives for large sales forces is well understood. A mobile application can streamline the process of communicating compensation and progress against goals back to the field. This helps keep representatives motivated and ensures they are spending their time on revenue-producing activity.

Access to the Company Intranet

Mobile workers need to access internal corporate data and documents frequently. Phone lists, vacation requests, expense report forms, travel policy, product information, and other common documents are used daily. Placing a local copy of a company's intranet information on a mobile device helps keep bandwidth costs down, while improving accessibility and providing more

Applying Mobility in the Workforce

rapid access to company procedure. This also allows for offline access to company forms where a network connection is unavailable.

Profiling Tools

The use of profiling tools allows the sales staff to view detailed information about clients and other influencers. Personal interest, purchasing patterns, and responsiveness to past promotions can all be tracked. This allows the company representative to tailor each sales call and personalize the experience for increased effectiveness.

Sales Analysis and Planning

Financial sales operations are flooded with data, both self-created and from third-party data sources. Sales representatives need powerful analytics to help them sift through the information, and develop strategies and plans for boosting sales in their territories. Interactive analysis and planning tools save time and help ensure best practices.

Interactive Selling Tools

Increasingly, sales representatives use slick multimedia interactive selling tools to structure their interactions with clients and provide a richer experience. The materials change frequently and can be automatically updated without significant cost.

Mortgage Origination

Real estate transactions are another data-intensive financial service greatly enhanced with the use of online forms and documents. Mortgage brokers will be able to expedite the loan pre-qualification and approval processes, help-

ing both buyers and sellers get fast and accurate information.

Mobile PIM (Personal Information Manager)

Communications such as flight and hotel confirmation numbers, cancelled appointments, and client emergencies are vital to the remote worker and may need swift action. Workers can stay more productive on the road with mobile access to up-to-date calendar and contact data.

Risk Management

Effective risk managers need to be able to provide swift assessments and deliver detailed recommendations based on those assessments. Automating the processes of risk management—including identifying, analysing, planning, tracking, controlling, and communicating—ultimately leads to less risk exposure for the client.

CONCLUSION

We are clearly at a time of change in the way that we service our customers. Companies around the world see value in a mobile workforce for a variety of reasons. According to The Bulletin's Mark Phibbs (2003), the main drivers will be:

- improvements in staff productivity,
- retaining quality people,
- providing better service to customers with rapid and appropriate response,
- providing tools to cross-promote products, and
- increasing field agents' contact hours with their customers.

These are only some of the reasons tempting companies into adopting a mobile strategy.

There are many factors that will affect how these solutions will be deployed. According to Intel's (2003) *unwired* advertising blitz that the growth of Wi-Fi hot spots will be the tipping point tip for the market, it is only a matter of time before we are all members of a mobile workforce. It would appear that perhaps we are now at, or at least close to the tipping point in terms of mass market appeal. The market seems to be moving away from Intel's desired *hotspot* solution and instead investigating deploying applications based on mobile telephone carriers' 2.5G and 3G technologies. GPRS has already gained mass market appeal, and 3G promises more of the same with much richer and more useful content.

There are hurdles when considering how to build a mobile workforce, but all of these seem to be manageable. Strong business cases for mobile access to enterprise systems are very likely to fuel investment in this space (Deshpande & Gilbert, 2002).

The transition to a mobile workforce need not be a painful exercise (Euronet, 2000). Following a set of guidelines, decision makers can make the transformation with ease. Some of the areas that do need consideration are to:

- build a mobile strategy on business use,
- make security a priority,
- have a framework to manage the devices deployed, and
- match the business need with the mobile coverage.

Instead of rolling out a mobile solution "just because it's cool", management should decide on a strategy that determines who will use it and how will it affect the business, what information the field personnel need, and how they could

access it remotely. Most importantly what is the likely return on investment, and is it a worthwhile proposition?

Security is perhaps the largest inhibitor affecting the decision of a company to invest in wireless technology (Gallagher, 2004). Any mobile solution must incorporate encrypting data, both as it travels in the air and when it is stored on the hard drive. Unscrupulous people will *sniff out* wireless data, and provisions should be put in place to ensure that the data is useless to them. In addition, strong authentication measures will help prevent leaking of valuable information.

The cost of a wireless solution will also depend on the variety of devices within the organisation. Similar products should be sourced to gain better discounts and lessen the technical support and training impact (Egan, 2003).

The choice of device should also be limited to the mobile provider's coverage and capacity available. There is no sense in arming a mobile workforce with tablet PCs so they can download large documents only to find the connection speed is too slow to be beneficial.

The hardware is already available, and GPRS mobile infrastructure is now well and truly in place. It can only be a matter of time before the mobile workforce is a common feature of mainstream business.

REFERENCES

- Bailey, R., Buist, C., & Vile, D. (2003). *Corporate wireless data in Europe*. Windsor, UK: QNB Intelligence.
- Broersma, M. (2004). Planes and trains: Wireless on the move. *Techworld Magazine*. Retrieved April 20, 2004, from <http://www.techworld.com/mobility/features/index.cfm?FeatureID=512>

Applying Mobility in the Workforce

- Computer Strategic Innovation. (1998). *Developing a mobile strategy—Preparing for the next discontinuous change*. London: Computer Sciences Corporation.
- Deshpande, N., & Gilbert, J. (2002). *GPRS: How does it work and how good is it?* Retrieved from <http://www.intel.com/update/departments/wireless/wi10021.pdf>
- Egan, B. (2003). *Making the case for wireless mobility investment*. North Providence, RI: Mobile Competency.
- Euronet Worldwide. (2000). *Keeping a wireless world connected*. Retrieved from http://www.euronetworldwide.com/solutions/biz_lines/software/mobile_banking.asp
- Gallagher, H. (2004). *Security still reigns as wireless's weakest link*. Retrieved from <http://www.macnewsworld.com/story/32874.html>
- Gessel, B. (2001). *Editorial comment: Should the telecom industry rethink architecture of its wireless networks?* Retrieved from <http://www.mobileinfo.com/Editorial/August15.htm>
- Gold, A. (2005). *Managing mobility in the enterprise*. Northborough, MA: A. J. Gold Associates.
- Hildebrand, C. (2004). *Steps to success: Managing mobile technology*. Retrieved from http://searchcio.techtarget.com/originalContent/0,289142,sid19_gci959674,00.html
- Intel Corporation. (2004). *Wireless LAN technology*. Retrieved from <http://www.intel.com/business/bss/infrastructure/wireless/solutions/technology.htm>
- Intel Corporation. (2004). *Wireless security best practice*. Retrieved from http://www.intel.com/business/bss/infrastructure/wireless/security/best_practices
- Intellisync. (2003). *Mobile strategies for financial services*. San Jose, CA: Intellisync Corporation.
- Kuykendall, L. (2004). *Retail-minded banks shift focus to traffic service*. New York: American Banker.
- Phibbs, M. (2003, August). Mobile workforce. *The Bulletin Magazine*. Retrieved from http://www.chamber.org.hk/info/the_bulletin/aug2003/it.asp
- Ramsaran, C. (2004). *Tripping over wireless technology*. Retrieved from <http://www.banktech.com/news/showArticle.jhtml?articleID=18600398&pgno=3>
- Softlab Ltd. (2004). *The mobile workforce*. Retrieved from <http://www.softlab.co.uk/fm/142/WPThemobileworkforce.pdf>
- Sweeney Research. (2002). *Mobility and mistrust*. Melbourne, Australia: Sweeney Research.
- Synchrologic. (2003). *The CIO's guide to wireless*. Atlanta, GA: Synchrologic.
- Urich, K. (2002). *Eight steps to going mobile*. Everett, WA: Intermec Technologies Corporation.

Chapter LVI

Applying Mobile Technologies to Banking Business Processes

Dinesh Arunatileka

University of Western Sydney, Australia

ABSTRACT

This chapter discusses the impact of mobile technologies on service delivery processes in a banking environment. Advances in mobile technologies have opened up numerous possibilities for businesses to expand their reach beyond the traditional Internet-based connectivity and, at the same time, have created unique challenges. Security concerns, as well as hurdles of delivering mobile services “anywhere and anytime” using current mobile devices with their limitations of bandwidth, screen size and battery life are examples of such challenges. Banks are typically affected by these advances as a major part of their business deals with providing services that can benefit immensely by adoption of mobile technologies. As an example case study, this chapter investigates some business processes of a leading Australian bank in the context of application of mobile technologies.

INTRODUCTION

Electronic commerce has become a dynamic force that has changed the way businesses operate on a global scale (Shi & Wright, 2003). Due to increased globalization, individuals, organizations, and governance frameworks have an increasing dependence on communication technologies. The Australian Communication

Authority envisions that ubiquity is the “best possible outcome” in terms of the future of business and economy in the country. This ubiquity is based on the elements of technology, market dynamics, users, and rules and guidelines (ACA, 2005). All business organizations in this global context are forced to look at this “best possible outcome” in order to stay competitive. This gives rise to several research

questions in the areas of business practices as well as workflow management, and affects the individual and collective social behaviour (Mylonopoulos & Doukidis, 2003). The research areas also focus on the mobile technologies and their application to businesses, with particular emphasis on the method and manner in which services can be delivered using mobile processes. Mobile processes are business processes that are executed with the use of mobile devices such as PDAs (personal digital assistants), mobile phones, or mobile-enabled laptop computers. Thus, mobility, which is the ability to move freely while performing regular business activities, has become an extremely crucial aspect of today's business processes. Furthermore, as per Archer (2004), in order to incorporate mobility, business processes also have to undergo substantial changes themselves to make it essential that the changes are researched and experimented into.

Internet Usage in the Banking Sector

Banks, as primary institutions of service-oriented business, have increasingly leaned towards e-commerce-based operations. Emerging mobile technologies offer "anytime, anywhere" type of banking that results in better customer orientation and provides personalization of services to the customer. The concept of banking using handheld devices, such as PDAs or other mobile devices, is becoming popular as it enhances the Internet connectivity to the fingertips of the customer (Unnithan & Swatman, 2002). The Internet has also provided opportunities for service providers such as PayPal, an online payment processing company founded in 1999, to offer more cost-effective payment-related services similar to banking services to its customers. PayPal, after a mere four years of operation, has become the

most used payment system for clearing auction transactions on eBay (Schneider, 2004), competing directly with the traditional banks. Banks thus face a major challenge and are forced to effect substantial cost reductions in order to be more competitive and offer cost-effective services to its customers. Banks aggressively push their customers to use electronic means for most of their banking, as these electronic transactions are far cheaper as compared to over-the-counter or ATM (automated teller machine) transactions. According to a recent study in the U.S., a teller transaction costs the bank U\$1.07. as opposed to a telephone transaction costing 54 cents, an ATM transaction costing 27 cents, a software-based PC transaction costing 1.5 cents, and an Internet-based transaction costing a mere 1 cent (Money Central, 2005). Mobile devices enable secure and convenient use of e-banking, payments, brokerage, and other types of transactions which are part and parcel of the banking sector (Herzberg, 2003). Another study reveals that among the Internet-based banking users, there is a positive tendency to use mobile devices to do banking transactions (Coutts, 2002). Hence, factors determining the success or failure of the mobile business and how the corresponding mobile systems and applications are designed, in order to provide banks with cost-effective, flexible, and customer-oriented business processes, are of interest to the banking community. The fact that today's banking customers are more educated, along with increasing demand for state-of-the-art services, also add pressure and push the banks towards mobile technologies.

Global Banking Industry

The educated and technology-savvy customers demanding better service and state-of-the-art technology is a global phenomenon in

the banking industry. For example, the banking industry in Europe is undergoing substantial changes as it looks to reduce costs and enhance the utility for customers through new technology. European banks are focused more on their core capabilities while exploring different sourcing options for non-core capabilities. They are disaggregating their value chain into independently operable functional units (Homann, Rill, & Wimmer, 2004). Furthermore, as communication capabilities reach higher levels of performance and reliability, these functional units are combined across corporate borders, providing valuable e-collaborations and flexibility for the organization. The industry sectors are changing the way they do business by using many different collaborations with customers (B2C), service providers (B2B), funding organizations (e-payments), government (B2G), and even competitors (B2B) (Arunatileka & Arunatileka, 2003). The emerging mobile financial applications including both mobile payments and banking services are also being investigated, showing how new financial services could be deployed in mobile networks and also identifying key players in the mobile financing value chain (Mallat, Rossi, & Tuunainen, 2004). Mobile customer relationship management is another area that would personalise business processes, adding value to organisations (Unhelkar & Arunatileka, 2003).

This chapter specifically investigates the implications of service delivery using mobile technologies. Since the author of this chapter has been researching within a well-known Australian bank (name is withheld due to confidentiality issues), this chapter is based on the service-delivery challenges related to incorporation of mobile technologies within the banking environment.

STUDY OF TODAY'S BANKING NEEDS

This section starts with a brief introduction of banking requirements studied in a leading bank (referred to merely as 'the bank') in Sydney. The bank was seeking to create a policy to introduce mobile technology to its banking processes and staff. This policy for mobile services (services using mobile processes) is meant to evolve from a broad framework of existing policies defined for the operations of the entire bank that also encapsulated its values and objectives. The vision of the bank is to inculcate three great values—namely, *teamwork*, *integrity*, and *performance*. The vision drives the organisation purposefully towards its objectives based on the four foundations of *staff*, *customers*, *corporate responsibility*, and *shareholders* (internal documents of the bank). The bank comprises different business units divided based on functionality for management purposes. These various units have trained staff in different disciplines. For example, the financial markets division would have trained investment advisors, the institutional banking division comprising trained relationship managers and customer care personnel, the retail banking division having trained home loan advisors and customer care personnel, whereas the human resources division trained human resource and training personnel. Thus, the bank consists of a multi-disciplined, heterogeneous workforce. The processes and the work methods of separate units could be very different from each other as well. For instance, the corporate and institutional banking divisions work internationally, thus having a 24-hour operation, whereas the retail banking is more likely to be an office hour operation subject to few exceptions. As mobility options that could increase productivity in corporate banking, PDAs

could be programmed to notify users of any new e-mails and SMS messages where the employees are notified immediately, and necessary action could be taken depending on the situation. This enables the employees to have more time with their families while still attending to urgent business. In retail banking, loan officers could use mobiles to be more competitive in the field. As parts of an organization, these various divisions have to work together as one entity in achieving the vision and goals of the entire organization.

The mobility policy, once accepted and incorporated, has to address the purpose and objectives underlined by the top management in the facilitation of better service delivery providing higher value to customers. This should fulfil most of the outlined areas by the management by facilitating the staff on better access, wider responsibility, and better tools, motivating them on better service delivery, achieving customer satisfaction, which in turn would fulfil the corporate objectives.

Business Units under Study

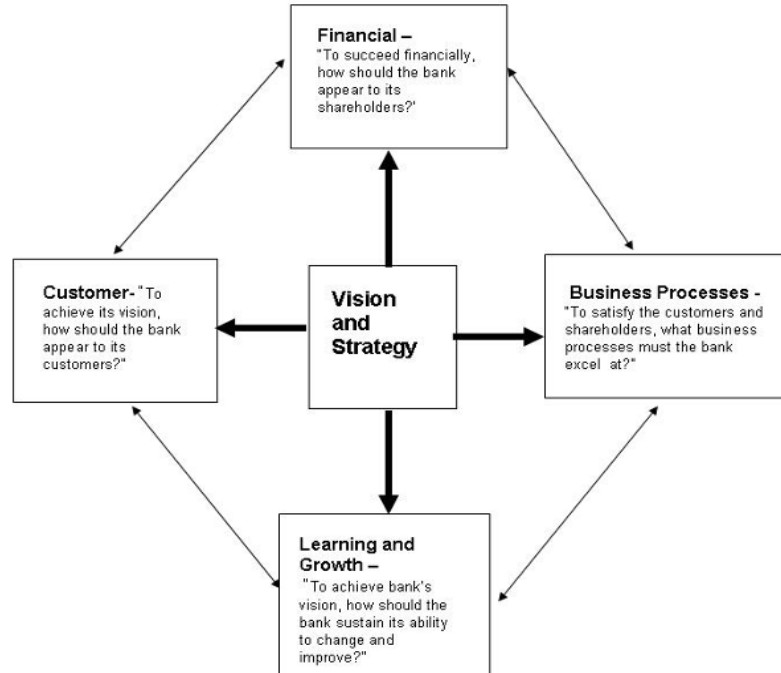
There were four business functional units identified by the bank as the initial study areas for introduction of mobile technologies, namely: financial markets, institutional banking, retail banking, and small investor operations. Although the last two units have a high volume of transactions, the first two functional units were given priority by the bank due to their high-value transactions and the need for change by the employees and the customers of these units. Furthermore, financial markets and institutional banking also had a pressing need to change their existing processes due to some existing limitations in their operations. Before the effect of mobility is investigated, a brief summary of how these two units work is described here.

- **Financial Markets:** This is a highly specialised unit, which brings the bank high revenues and very high profitability. Although the number of customers may be lower, the revenues are very high and revenue per customer is very high, resulting in these customers demanding and warranting individual attention. The concerned managers would be international business managers and state managers who are handling time-sensitive corporate accounts. These managers travel a lot, meeting customers and looking after their specific interests.
- **Institutional Banking:** This is also a very highly specialised unit where institutional banking managers generally go to their client organisations in order to serve their needs. Although there are less volumes of transactions, the values could be very high, bringing very high profitability to the bank. Most accounts would belong to large business organisations having diversified needs. Mobile access to the systems would make it much easier for these managers to be in touch with the latest communications, rates, and so forth, which would be very essential in serving business customers in a highly competitive market.

BACKGROUND TO RESEARCH IN THE BANK

The background reasons for researching into the bank's processes in the context of mobility were based on the long-term expectations of the bank. Usually, the expectations of an organisation are summarized in the mission and vision statements. The bank had very specific values and purpose, spelled out in its mission and vision statements. The bank also used the

Figure 1. A balanced score card for the bank



concept of a balanced score card in order to maintain the policy balance and drive its strategies. The balance score card concept (Balanced Scorecard, 2005) is built on four perspectives—financial, internal business processes, customer, and learning and growth, with the vision and strategy in the centre of it.

The bank envisions and makes strategies in the business with respect to these four perspectives. The four perspectives were adapted from www.balancescorecard.org, which describes the perspectives with respect to the organization. All four perspectives are so tightly entwined that one leads to the other. Financials is the starting point, which is the very core of the objectives of a business, simply to pass a value to its owners/shareholders. However, since the banking industry is highly competitive, the customers should be well looked after in order to retain them. Business processes, learning and growth, and customer perspectives all contribute to make the bank more customer oriented. The main focus in this chapter is aimed at

business process perspective, which refers to the business processes of the organisation. The measurement on this perspective allows the managers to know how well the business is running in terms of whether the products and services conform to customer requirements (the mission of the organisation). In addition to strategic management processes, two kinds of processes could be identified—mission-oriented processes and support processes (Balanced Scorecard, 2005). The applications of mobile technologies are looked at in all these processes.

The bank conceptually was looking at two major areas of mobility: mobility at the workplace and mobility in the delivery of service to customers. The mobility at the workplace is in line with a more futuristic plan for a new headquarters building with a smart office where the employees could work anywhere within the building. Mobility in delivery of service is more with the current business processes and how they could be improved to offer better service

delivery to customers. In-depth knowledge of the existing work processes, coupled with a detailed plan to transform the existing organization into a mobile enabled (m-enabled) organization without disrupting the day-to-day functions of the organization, is one outcome expected of this exercise. Technology has made progress from earlier setbacks, but the methodologies to fully implement the existing technology into the business operations have to be done carefully in a systematic manner.

The initial expectations of the bank were to look at the business unit of financial markets. The decision was financial and also partly need driven since the financial markets were a big earner and the managers in the unit were very keen to move forward with new technology. Once the focus area was identified, the focus was concentrated on the fundamental questions.

TRANSFORMATION OF BUSINESS PROCESSES

The financial markets area was selected as the first unit for this study since the most pressing demand for change was persistent in this unit. There were several research questions arising with this selection. They were:

- How does the service-based industry change with application of mobile technologies? A generic question focused on the service-based industry as a whole investigating the possibilities to improve efficiency while cutting down on long-term costs.
- What are the changes happening in the banking industry? This is an industry-specific question to look at what is happening elsewhere in the banking industry which is relevant to the current timeframe.

- What should be the bank's response to the change in the service industry and specifically in the banking industry? This probing question is looking inward into the bank's own processes critically to decide how change could be facilitated to improve on the internal processes.
- What is the expected impact after mobile technology is introduced to a selected business unit in the bank? A specific question arising from mobile technology being introduced to the unit which is expected to create a positive impact.
- What would be the direct impact on the customers once the new technology is fully implemented? A probing question to understand the impact on the customers once the change is made.
- Would there be any anticipated problems during the changeover? A question to understand the management of change from the existing to the mobile-enabled organization. It is also important to measure this change in terms of cost factors, time factors, customer satisfaction factors, security issues, and other such factors which the bank thinks important to consider.

Let us look at transforming a simple business process, like the checking of an account balance by a customer as an example. The bank would concentrate on the cost of the process, the satisfaction of the customer in the process, assurance of the process in delivering the right information, security issues in providing the service, and timeliness of the information from the customer's point of view.

If the process is m-transformed from checking the balance through the Internet or at an ATM to use of a mobile phone, how will such a transition be measured in terms of this process? In order to understand the m-transition of the

Figure 2. A business process transformation

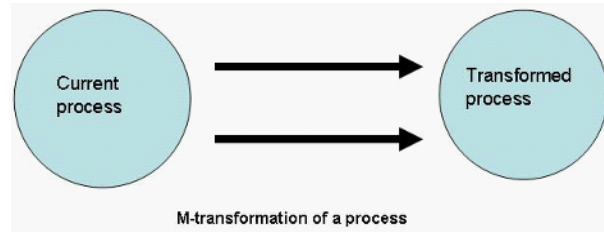
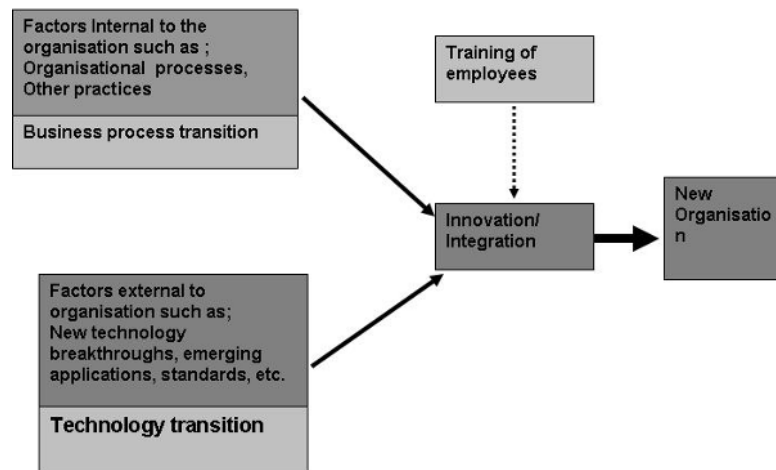


Figure 3. A mobile transition roadmap



process, it is essential that the process is understood and examined critically. The first step in such a critical evaluation is the creation of a mobile transition roadmap. Such a roadmap is proposed in Figure 3, which shows m-transition applied to a bank focusing on the overall picture while the transition is in progress.

The mobile transition roadmap will capture all the areas that have to be considered in m-transforming an organization in order to become a mobile-enabled organization. The concept of the roadmap for mobile transformation, as shown in Figure 3, has been adopted and evolved from the electronic transformation roadmap (Ginige, Murugesan, & Kazanis, 2001). The mobile transition roadmap is further analysed and investigated using different perspectives in Figure 8, later in the chapter.

The Business Processes in the Financial Markets Division

The Fund Managers and International Business Managers (FM/IBM) are called to service customers at anytime of the day since financial markets are a global business. Different market segments would be working in different time zones. It is important be on top, knowing what is happening all over, all the way from Sydney to New York through the other giants such as Japan, Hong Kong, and Europe. Therefore the managers working in this unit should be dynamically connected. This is one area where mobile technologies could be used very effectively to enhance productivity. The role of an FM/IBM in particular involves visiting clients

and understanding their requirements in international business from telegraphic transfers to structuring major import/export deals, to financing solutions for the funding of these transactions to mitigate risks associated.

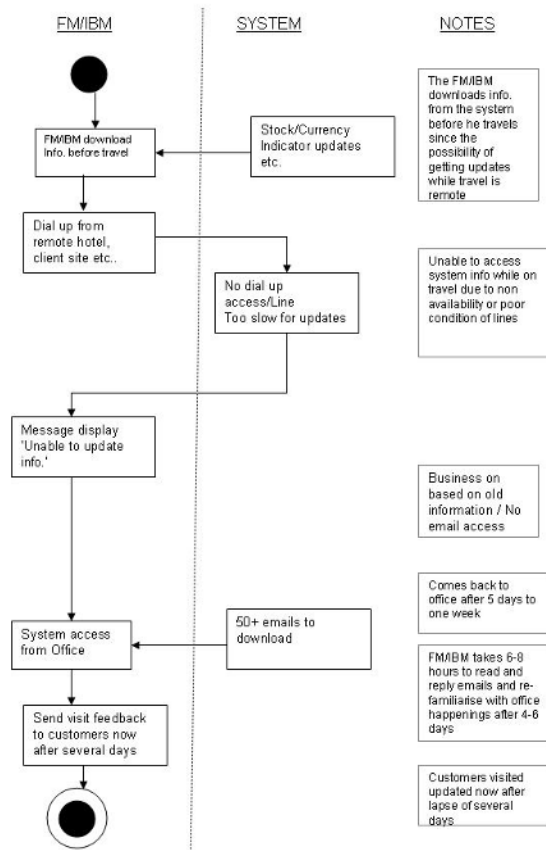
At present, laptops are being used which contain the data and programs to run the business that is provided online through the system. With the very mobile nature of the worker, being in the field for long periods visiting clients, the access to the system could pose problems. Dial-up access could be time consuming and frustrating. Response time would be one of the major factors for corporate customers in quantifying the service levels of the bank. The customers may be of a captive nature in the short term, and the exit may take some time due to contractual obligations. However, high ser-

vice levels could keep the corporate customers on a long-term basis in line with the bank's objective of "achieving at least a 5% increase in agreed customer satisfaction measures." Thus, the retention and growth would be of great interest to the bank. Expensive acquisition is also valuable in this category of high value customers.

Example 1—Transition of Travel Process of FM/IBM

Travel process is discussed herein where the mobile-transformation was suggested. The process of travel in the financial markets division is undertaken by an FM/IBM. These managers travel often to meet customers all over Australia.

Figure 4. Activity diagram, FM/IBM travel—current process (before m-transition)



Two activity diagrams are drawn to show the travel process before and after mobile transition. The activity diagram which is a tool from the Unified Modeling Language (UML) is used herein as it is a typical analysis tool. Figure 4 depicts the picture before m-transition and the activities of the FM/IBM involved in the process.

In Figure 4, the current process for travel for the FM/IBM has been depicted. The FM/IBM visits the customers but does not have access to the bank's system most of the time due to bad lines and low line speeds. Therefore the FM/IBM would visit customers and be travelling for about five days every two months with no access to the bank's system most of the time. During this time, the FM/IBM is completely cut off from the news in the bank and e-mails from their own customers as well. Moreover, the customers visited during the trip will not get any feedback for a considerable time until the FM/IBM has access to the bank systems. When he or she finishes travelling and gets back to office, he or she would have to take six hours (approximated, based on current estimates) to read and respond to e-mails before starting on his or her other work such as making proposals for the customers visited.

This process of travel is modelled using an activity diagram (which is like a flowchart, and is derived from www.omg.org) for easy comparison. Note that the customer, the most important person in the process, does not appear as an actor in the current process, since the FM/IBM has to wait until he or she gets back to the office to respond to the customers. Thus, the customer is not in the current process at all active.

Figure 5 depicts the scenario once the existing processes from Figure 4 are transformed, with mobile transition taking place. Each FM/IBM would each his or her laptop and could dial into the bank systems. The information would be initially text based so that speed problems could be overcome at initial stages until mobile

technology is mature enough to deliver high bandwidth without problems.

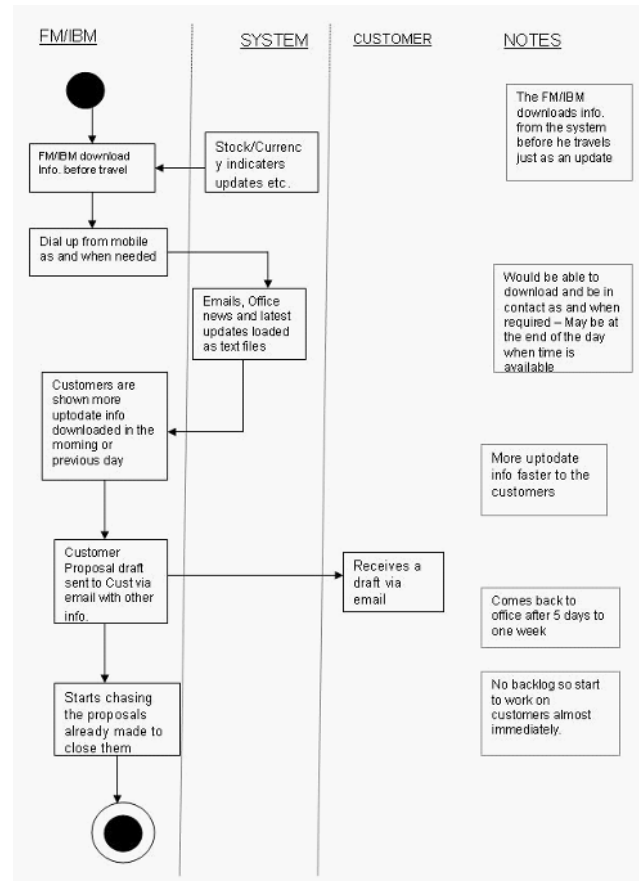
The FM/IBM could download e-mails and relevant figures every night or every morning as and when he or she has free time before or after visiting customers while travelling. This would enable him or her to be in contact with the bank regularly. Customers could get draft proposals via e-mail since the FM/IBM could do the proposals quickly after visiting customers without having to wait until he or she gets back to the bank. It is also time saving since he or she is updated regularly and does not have to spend six long hours reading and responding to e-mails as per current estimates, since all that has already been done during his or her free time while travelling. Also note that the customer is an actor in this process. The customer gets the feedback while on travel.

Thus, mobile transition should save a lot of time for the managers in downloading and reading e-mails, and also should keep them in line with the current rates in the very volatile financial markets area. Also the entire sales process has been expedited, with the customers getting the draft version of their contracts very early. Thus the sales process would be shortened by several days, which could bring substantial income. Customers should also perceive this state-of-the-art process positively, which appears to be very active in comparison to the existing process.

Example 2—Transition of Customer Meeting Process of FM/IBM

The second process discussed herein is that of customer meetings of the FM/IBM. This is also very important for the bank since these meetings could bring very high-value business. Most of these meetings would be happening during the travel process. Therefore, the access to bank computers and systems will only be avail-

Figure 5. Activity diagram, FM/IBM travel—new process (after m-transition)



able through dial up. Speeding up the process would be profitable to the bank on the one hand and also would give them an advantage on customer service on the other.

Customer Meeting Process before M-Transition

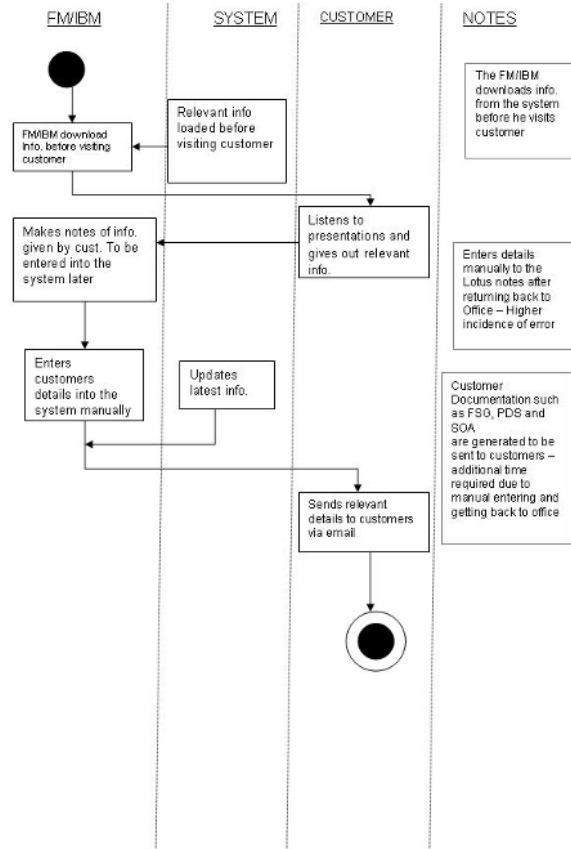
Figure 6 shows the current process of customer meeting for the FM/IBM. The FM/IBM visits the customers, but does not have any forms in his or her laptop to enter any data at the customer's site. The FM/IBM must get information and then come back to the office and manually enter this data to generate relevant customer documentation. Thus, there is a considerable time gap until the customer receives

final documentation. This could be subject to errors, and due to manual entry without verification, time is needed to get customer feedback after the entry.

Customer Meeting Process—After M-Transition

In Figure 7, the current customer meeting process has been changed with the introduction of mobile technology. The new process anticipated after the m-transition has taken place is shown. The FM/IBM would have his or her laptop preloaded with forms to enter customer data, which could later be loaded into Lotus Notes. Data is entered and verified at the customer's site, saving considerable time. The

Figure 6. Activity diagram, FM/IBM customer meeting—current process (before m-transition)



customer could get a draft report until final report (uploaded and updated with latest data) reaches him.

Considerable time savings is apparent with m-transition taking place. This would be crucial in higher productivity for the FM/IBM, while customers perceive higher service levels as well as remain ahead of the competition.

THREE PERSPECTIVES IN M-TRANSITION

To build on Figure 3, where transition of business processes and technology were merged in order to look at new business processes, Figure 8 describes the typical requirements on the

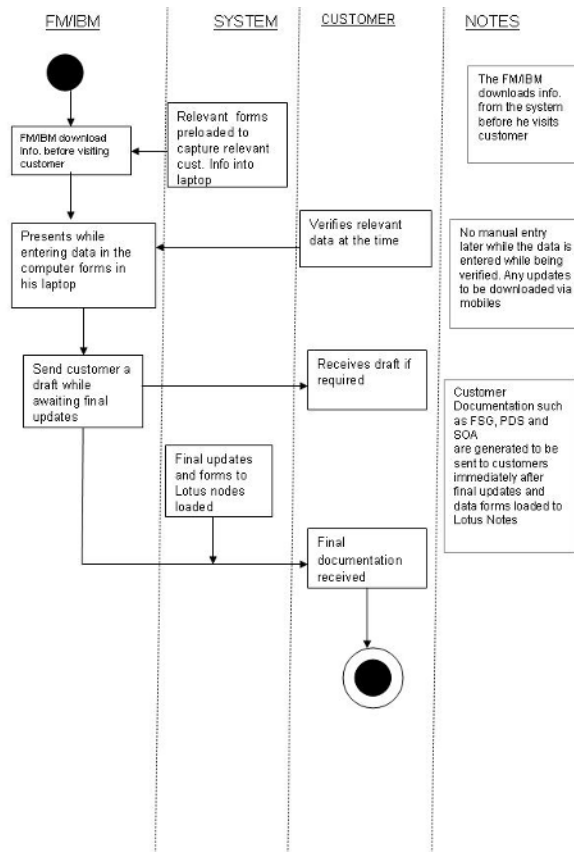
bank's side, generalizing all the processes taken into consideration. The diagram considers the processes kept intact, changed, and scrapped, and also looks at the entire change from technology, methodology, and sociology perspectives (Unhelkar, 2003).

These perspectives are useful in managing the entire organizational change due to m-transition.

Technology Perspective

The issues such as what applications should be used, how these applications should be integrated and the networking of these applications, the security issues in the new organization (the mobile transformed organization), and

Figure 7. Activity diagram, FM/IBM customer meeting—current process (after m-transition)



the devices that could be used to facilitate the employees in the service delivery aspects must be decided. In the FM/IBM travel example, the laptops and how to deliver connectivity while travelling has to be looked at from the technology perspective. The aspect of altering certain bank systems' fit enough for the mobiles to download them into a laptop computer has to be investigated. This technology is currently available.

Methodology Perspective

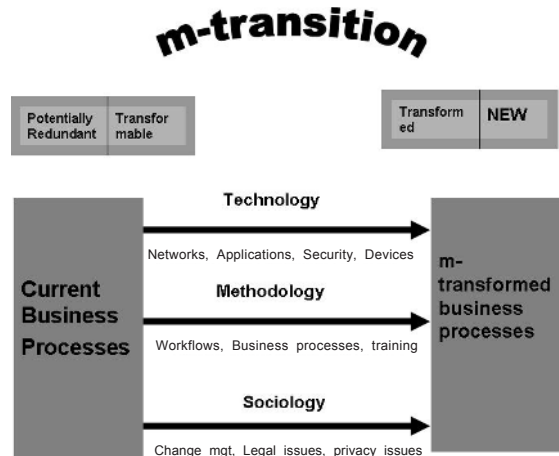
All the procedures that should be followed and adhered to in adopting the new business pro-

cesses and the approach to be followed in order to transform into the new organization should be discussed. It is also very important that the employees are trained on the new security measures, devices, and new business processes. In the travel example, the new processes followed in order to download e-mails, delivery of business proposals, and so forth are the considerations in the methodology perspective.

Sociology Perspective

Wider issues include the management of the entire change process along with any legal implications and privacy issues in the new

Figure 8. Mobile transition of business processes in technology, methodology, and sociology perspectives



organization. Providing training to employees in these areas is also of utmost importance. In the travel example, the training of the FM/IBM on the new processes, any legal issues falling therein in giving the managers the authority to provide preliminary proposals, and so on will fall within the sociology perspective.

Table 1 shows that considerable time savings could be achieved in m-transition. There are other measures as well that are significant to consider. The customers who are mostly corporate clients would like to get up-to-date

information. This falls within the customer perspective of the balanced score card of the bank.

Table 2 shows the comparison of the current and the proposed processes for customer meetings. Considerable time savings appear on delivery of proposals, and so forth. This also leads to better accuracies, as the data is entered at the customer site and verified then and there. Thus considerable monetary savings result due to sped-up process, and also additional revenue results due to customers being signed up earlier than the current system.

There would be additional costs for laptop and software upgrades in the proposed processes. However, benefits would be much more compared to costs involved since large time savings and additional revenue would compensate for one-time costs of upgrades. There will also be intangible benefits such as better customer satisfaction due to timely delivery and better employee satisfaction due to saving of considerable time, which could be used to make more customer visits.

The main concentration was on the business process perspective of the balanced score card. However, the correct implementation of this perspective would lead to improvements on the customer perspective leading to learning and growth and financial perspectives in a rolling effect.

Table 1. Comparison of the current processes with the proposed processes with regard to travel of the FM/IBM

Activity/Attribute	Current Process	Proposed Process
1. Downloading information before travel	Stock/currency indicator updates at office	Stock/currency indicator updates at office
2. Dial up for updates during travel	Too slow and difficult for updates	Mobile dial up for updates as frequently as necessary
3. Liaison with office	Infrequent/almost nil due to bad lines, etc.	Always in touch with frequent periodic updates
4. E-mail contacts	Almost nil	Periodic updates every evening with office
5. Feedback to the customers visited	Several days since the visit	The proposal could be delivered within a day
6. Timeliness of information on hand	Could be several days old	Only few hours since last update

Table 2. Comparison of the current processes with the proposed processes with regard to customer meeting of the FM/IBM

Activity/Attribute	Current Process	Proposed Process
1. Downloading information before visiting customer	Relevant forms are downloaded at office	Relevant form downloaded at office is updated with current information just before visit
2. Verification of customer information	Makes notes to do corrections later while entering	Updates and verifies data at current time before providing a draft to customer at his/her site
3. Re-verification of customer information	May be needed to be done via e-mail due to manual entering	Not required since it is already done
4. Liaison with office	Infrequent/almost nil due to bad lines, etc.	Always in touch with frequent periodic updates
5. E-mail contacts	Almost nil	Periodic updates every evening with office
6. Feedback to the customers visited	Several days since the visit to enter the information and verifications	The proposal could be delivered within a day
7. Timeliness of information on hand	Could be several days old	Only few hours since last update

CONCLUSION AND FUTURE DIRECTIONS

The transformation of the travel process and customer visit process of fund managers and international business managers introducing mobile technology into the bank created a situation wherein the bank stood to gain significantly in terms of savings of time and money. The mobile transition also highlighted potential intangible benefits such as better customer focus, timely proposals boosting customer satisfaction, and significant time savings, leaving the managers with more time to focus on their customers. The outlook created by the intangible benefits should also lead to gain more customers. However, a systematic approach is suggested, introducing the change gradually while educating the employees to be aware of the change. The customers are also being included in this process of gradual change towards an m-enabled organization.

Similar processes in the financial markets division and other divisions are currently under investigation with a view for further m-transformation. Once the other significant processes

have also been transformed across all units, to use mobile technology effectively, the chances of the bank achieving a perfect balanced score card will be significantly enhanced.

REFERENCES

- Archer, N. (2004). The business case for employee mobility support. In *Proceedings of the IADIS International Conference in E-Commerce*, Lisbon, Portugal.
- Arunatileka, S., & Arunatileka, D. (2003, December). E-transformation as a strategic tool for SMEs in developing countries. In *Proceedings of the 1st International Conferences on E-Governance*, New Delhi, India.
- Australian Communications Authority. (2005). Vision 20/20: Future scenarios for the communications industry—implications for regulation. *Final Report*, (April).
- Balanced Scorecard. (2005). *What is the balanced scorecard?* Retrieved April 17, 2005, from <http://www.balance scorecard.org>

- Coutts, P. (2002). *Banking on the move*. White Paper, Communications Research Forum.
- Ginige, A., Murugesan, S., & Kazanis, P. (2001, May). A roadmap for successfully transforming SMEs in to e-businesses. *Cutter IT Journal*, 14.
- Herzberg, A. (2003). Payments and banking with mobile personal devices. *Communications of the ACM*, 46(5), 53-58.
- Homann, U., Rill, M., & Wimmer, A. (2004). Flexible value structures in banking. *Communications of the ACM*, 47(5), 34-36.
- Mallat, N., Rossi, M., & Tuunainen, V. K. (2004). Mobile banking services. *Communications of the ACM*, 47(5), 42-46.
- Money Central. (2005). *MsMoney.com—online banking—online fees*. Retrieved April 16, 2005, from <http://www.moneycentral.com>
- Mylonopoulos, N. A., & Doukidis, G. I. (2003). Mobile business: Technological pluralism, social assimilation and growth. *International Journal of Electronic Commerce*, 8(1), 5-21.
- Schneider, G. P. (2004). *Electronic commerce: The second wave* (5th ed.). Thomson Course Technology.
- Schwiderski-Grosche, S., & Knospe, H. (2002). Secure mobile commerce. *Electronics & Communication Engineering Journal*, 14, 228-238.
- Shi, X., & Wright, P. C. (2003). E-commercializing business operations. *Communications of the ACM*, 46(2), 83-87.
- Unhelkar, B. (2003). *Process quality assurance of UML-based projects*. Reading, MA: Addison-Wesley.
- Unhelkar, B., & Arunatileka, D. (2003, December). Mobile technologies, providing new possibilities in customer relationship management. In *Proceedings of 5th International Information Technology Conference*, Colombo, Sri Lanka (pp. 23-31).
- Unnithan, C. R., & Swatman, P. M. C. (2002). Online banking vs. brick and mortar—or a hybrid model? A preliminary investigation of Australian and Indian banks. In *Proceedings of the 7th COLLECTeR Conference*, Melbourne, Australia.

Chapter LVII

Mobile GIS—Challenges and Solutions

Pramod Sharma

The University of Queensland, Australia

Devon Nugent

The University of Queensland, Australia

ABSTRACT

This chapter focuses on Mobile GIS (MGIS), which uses wireless networks and small screen mobile devices (such as PDAs and smartphones) to collect or deliver real time, location specific information and services. Such services can be divided into field and consumer (location based services) GIS applications. The use of wireless networks and small screen devices, introduce a series of challenges, not faced by desktop or wired internet GIS applications. This chapter discusses the challenges faced by mobile GIS (e.g. small screen, bandwidth, positioning accuracy, interoperability, etc.) and the various means of overcoming these problems, including the rapid advances in relevant technologies. Despite the challenges, many efficient and effective Mobile GIS applications have been developed, offering a glimpse of the potential market.

INTRODUCTION

A geographic information system (GIS) is a computer-based system designed for the collection, storage, analysis, and visualisation of geographic data. Geographic data includes geographic location as an important attribute. The technology of GIS has undergone rapid development over the past three decades and in the process has transformed itself from mainframe-

based systems to Internet-based distributed systems operating on a variety of hardware platforms (see Table 1). During this period, GIS applications have also changed from “the static compilations of the specialist to applications supporting the everyday lives of everyone, everywhere, all the time” (Smyth, 2000). The GIS hardware, software, and services industry was valued at over US\$7 billion in 1999 and growing at over 10% per annum—estimated to be over

US\$11 billion in 2004 (Longley, Goodchild, Maguire, & Rhind, 2001, p. 13; Daratech, 2005).

Early GISs of the 1970s were static, stand-alone, proprietary systems, focusing on inventory applications (e.g., inventory of natural resources, transportation networks, or utilities infrastructure) and on the automation of existing tasks. The next phase in the evolution of GIS involved the use of the networked client-server model to access remote data servers, and more advanced analysis and modelling capabilities. These systems were still, however, closed, stand-alone systems. They were used to model soil erosion, predict flood risk, model power network outages, and so forth. The GISs of today, however, are open distributed systems utilising the wired and wireless Internet to access distributed GIS services, tools, and spatial information for real-time data management applications (e.g., emergency management systems, location-based services). Not so obvious over the period has been the change in emphasis from “GIS software” to an emphasis on “GIS functionality”—the latter not necessarily delivered via “GIS software”.

Traditional GISs are large project, departmental, or enterprise-wide PC- or mainframe-based applications, with full GIS functionality (e.g., natural resource inventory, urban management systems, utilities management systems). Such “legacy” applications continue alongside the newer types of Internet-based GISs—indeed, they are often the core component of the newer applications. Also, while the evolution of technology deserves analysis in its own right, it is the change in the user base (“market”) rather than the technology change that is influencing new developments—technology is merely the enabler. Thus Internet GISs tend to have more limited functionality, and can be accessed by clients without GIS software and with little GIS experience/knowledge. The needs of these users are easily met

by simple mapping output to simple queries: Where is x? Where is nearest x? How do I get there? Internet GIS applications do, however, allow a much wider range of people to gain access to GIS tools and data via the wired Internet (e.g., the usage of MapQuest and WhereIs type services).

This “simplification” or “democratisation” of GIS technology finds a natural home in Mobile GIS (MGIS). In MGIS, current technology limitations—both in terms of the wireless communications infrastructure as well as those related to small-screen mobile devices—introduce additional constraints to GIS functionality. MGISs are, however, very well suited to a wide range of field and consumer applications, which do not require a full GIS toolset, processing power, and so on. MGISs complement traditional GISs, extending some GIS functionality into the field (e.g., to collect and update the databases of enterprise GISs).

This chapter focuses on MGIS—where the service coverage is wireless based and the client platforms are small-screen mobile and wireless devices, such as laptop computers, tablet PCs, PDAs, and cellular phones. We exclude laptop computers from any further analysis as, apart from the wireless connection, they essentially mimic the functionality of desktop computers. The chapter begins with a discussion of what MGIS is and the rationale for its introduction. It then examines some common applications, the challenges faced by MGIS, and some of the solutions employed to overcome these problems. It will conclude with a look at the future directions of MGIS services.

MOBILE GIS—THE INNOVATION

What is MGIS? MGIS refers to the access and use of GIS data and functionality through mobile and wireless devices such as mobile laptop

Table 1. Evolution from static to Mobile GIS

Stand-Alone GIS 1970s-1980s	Early Network GIS Mid-1980s-1990s	Complex Network GIS Late 1990s -	Mobile GIS 2000s -
Large, project applications, GIS software, and data reside on fixed computers; full GIS functionality; proprietary data formats; closed, stand-alone systems; applications concentrate on the automation of existing tasks; static inventory applications (e.g., natural resource inventories, early urban management systems, tax assessment); computer cartography	Data can be accessed on a server via LAN or WAN; allows centralised storage and remote access to data; software must be installed on fixed computers; full GIS functionality; proprietary data formats; closed, stand-alone systems; more advanced analysis and modelling applications (e.g., modelling flood risk, power network outages, transportation networks)	Distributed GIS; software, and data reside on one or more remote servers; clients can access GIS tools, services, and data through Web browsers; existing Internet applications are customised GIS applications with limited functionality (simple analysis and mapping); utilises open standards (e.g., OpenWMS); many desktop or professional GISs also allow access to remote data via the Web	Infrastructure features similar to complex network GIS; data and GIS software tend to reside on the server; some software can be installed on the mobile device (e.g., ArcPad); utilises wireless networks; a limited but specialised set of tools for fieldworkers and consumers; highly customised applications; utilises mobile devices; standards based; real-time data management applications (e.g., utilities maintenance, disaster management); location-based services

computers, tablet PCs, PDAs (such as Palm Pilots and pocket PC devices), and Web-accessible smartphones. They differ from traditional GISs and Internet GISs in that they utilise wireless networks and small-screen mobile devices. MGIS applications development requires a complete rethink of existing GIS models, although they are essentially a “repackaging” of Internet GISs for mobile devices, offering functionality similar to Internet GIS applications that utilise the thin client model (Peng & Tsou, 2003, p. 479). However, it is not necessary to have specialised MGIS software in order to offer such applications (e.g., such services can be provided using J2ME Location API, Oracle Spatial, and Oracle Locator). Major specialised MGIS software products include: ArcPAD from ESRI, MapXtend from MapInfo, IntelliWhere from Intergraph, and OnSite from Autodesk.¹

Why MGIS? As GIS has evolved from essentially static systems to a technology which

is available to “everyone, everywhere, all the time” (see Table 1), it is able to address needs which have resulted in the creation of MGIS. These needs include (Hassin, 2003):

- access to geodata in the field where it is often needed the most,
- ability to capture data in the field and in real time,
- ability to append positional information to data capture, and
- GIS functionality where it is often needed the most.

Given these needs, it is not surprising that the driving force behind the development of MGIS has been the needs of fieldworkers, as well as meeting the management’s objectives of “field force automation” (i.e., to improve the work process of field service or field sales employees). MGISs have revolutionised field data collection through the many advantages

they offer to fieldworkers and organisations: improved accuracy of field data collection and update; improved access to relevant and up-to-date data; improved efficiency in data collection, validation, and analysis; improved timeliness of the decision-making process; and so forth (ESBI Computing, 2005a; ESRI, 2005).

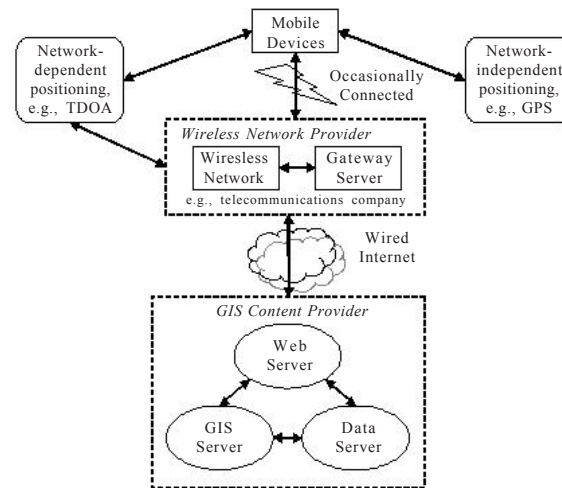
MGISs have also enabled users with small-screen mobile devices to access GIS applications and spatial information that was previously restricted to fixed desktop or workstation applications (even Internet GIS applications were only available for large-screen devices). The widespread use of mobile phones, especially smartphones, and the increasing use of PDAs for managing day-to-day activities have also led to the extension of MGIS applications for consumers. Indeed, for many users the main use of GIS services is when they are mobile (e.g., users of various directories, locator or navigation services).

What is the basic MGIS architecture? MGISs have unique requirements and considerations that set them apart from other types of GISs. MGISs require the following components:

- a mobile device that is capable of being detected by some position-determining technology,
- position-determining technology,
- a GIS content provider (Web server, GIS server, data server), and
- a wireless network provider (including gateway server).

They differ from wired Internet GISs in three respects: (i) the use of wireless networks to connect the mobile devices, (ii) the use of small-screen mobile devices, and (iii) a gateway server is required to interface the Web server and wireless network (Peng & Tsou, 2003, pp. 477-479). The first two components have major impacts on the applications that can

Figure 1. Components of an MGIS



be developed, as they impose a set of significant constraints. On the positive side, they do allow the development of a new range of services that were not previously possible.

The wireless network provides a means of uploading or downloading information to or from a server to the workers in the field. Positioning technologies, particularly GPSs, enable the recording of the location of the mobile device and so can be used to record the location of features and objects at a known level of accuracy. Hence, MGIS has demonstrated the capability for real-time data collection and delivery in the field for a wide range of purposes (see Table 2).

APPLICATION CASE STUDIES

The major users of MGISs are field workers (data collection + service delivery) and consumers who need real-time geographic information and services (Peng & Tsou, 2003, p. 492). Hence, applications can be divided into Field GISs (e.g., utilities maintenance, pavement management, responding to outages, re-

Mobile GIS

Table 2. An overview of potential MGIS services²

	Applications	Example
Field GIS	Data collection	Recording soil sample data
	Data update and validation	Updating information on crop conditions
	Utility and facility maintenance (fieldwork/repairs)	Pavement management; telecommunications repairs
	Asset inventory and maintenance	Collect information on assets and their condition, e.g., street lights
	Customer repair services	Managing telecommunications repair crews
	Site analysis/inspections	Recording the presence of fire ants
	Incident reporting	Reporting the location of a car accident
Location-Based Services	Tour guide	Pre-defined tour around Brisbane City
	Routing	Best route from the hotel to the Queensland Museum
	In-vehicle navigation	Best route from home to the CBD
	Traffic information	Updates on traffic accidents, congestion, and so on in the nearby area
	Fleet tracking	Tracking parcel delivery vehicles
	Emergency response	Transmits the location of an accident to emergency services; finds the best route to an accident site; displays real-time information on the spread of a bushfire
	Concierge services	Where is the nearest Italian restaurant?
	Location-sensitive billing	Subscribers charged according to their location
	Marketing	Information on specials at a specific retailer
	Child safety	Monitoring where children are and if they move outside certain approved areas
	Criminal (release) tracking	Ankle bracelet that home release prisoners wear

Table 3. Fieldwork GIS case study 1 (Geological Survey of Western Australia, 2005)

Application Area	Asset inventory and maintenance
Summary	The Geological Survey of Western Australia (GSWA) is using Mobile GIS and GPS to map the location of abandoned mine sites (mines that closed before 1990). The aim of the project was to locate and document abandoned mine sites, identify safety and environmental hazards, and assess their state of preservation.
Push/Pull Services	Collect data on the location and condition of abandoned mine sites
Geographic Coverage	Western Australia
Positioning	LinksPoint's GlobalPoint GPS receiver
Mobile Devices	Symbol PPT 2800 mobile computers; now use an integrated PDA/GPS
Platform	ArcPad
Communications Network	Downloaded to laptop in the field and database updated later
Comments	Allows for effective and accurate data collection, and improved productivity

Table 4. Fieldwork GIS case study 2 (North London Strategic Alliance, 2005)

Application Area 2	Incident reporting
Summary	North London Strategic Alliance (NLSA) is using Mobile GIS in order to combat crimes “such as graffiti, fly-tipping, anti-social behaviour and vehicle abandonment and to improve environmental conditions that create the opportunity for these activities.”
Push/Pull Services	Report the location and details of crimes in real time; allow immediate action to be taken
Geographic Coverage	North London, UK
Positioning	GPS
Mobile Devices	-
Platform	ArcPad, ArcSDE database hosted by ESRI; database can be accessed on a secure Web site hosted by ESRI (using ArcIMS)
Communications Network	GPRS
Comments	Mobile GIS is seen as “improving the end-to-end cycle of reporting an incident to taking corrective action.” It improves efficiency and productivity. The recording of such information also allows for spatial analysis, the identification of trends, and hotspots.

cording accident locations) and Consumer GISs or location-based services (LBSs). This classification highlights a fundamental divide in the MGIS industry, and as such it is also a useful framework for discussing almost any aspect of the industry.

The main advantages offered by MGISs for fieldwork are the ability to collect, enter data, validate, and manage geographic data in real time at the relevant location. The use of this technology eliminates delays in data collection, data entry, validation, and database updates (Peng & Tsou, 2003, p. 453). Hence, the immediate advantage for field workers is apparent. Fieldworkers can also utilise MGISs in order to navigate to clients or facilities needing repairs, to manage repair crews and jobs in real time, and to access information in order to carry out specific tasks such as telecommunications repairs (Peng & Tsou, 2003, pp. 493-495). The use of Mobile GIS improves the efficiency, accuracy, and productivity of fieldwork. At the same time the system also permits managers to manage their field workers. Consequently, it is easy for enterprises to justify their investment in MGIS.

The Consumer submarket of MGIS is essentially what is referred to as location-based services, although some LBS applications (e.g., vehicle tracking systems) do not sit comfortably under the Consumer MGIS label. Both terms refer to a suite of applications which utilise the location of a mobile device in order to deliver real-time, location-specific, value-added, personalised services. The worldwide location services market is expected to increase from US\$.7 billion in 2003 to US\$9.9 billion by 2010 (Giaglis, Kourouthanassis, & Tsamakos, 2002, p. 65). LBSs encompass a wide range of services from in-car navigation (e.g., Navman GPS) to mobile tourist guides, concierge services, emergency services, location-specific marketing, and “find a friend” services. These services can be categorised into ‘push’ and ‘pull’ services. Pull services are those services elicited by the mobile client (e.g., where is the nearest Chinese restaurant?). Push services are those initiated by the service providers (e.g., location-specific marketing). These services support the client on the move and would have to be monitored to avoid becoming a nuisance (e.g., a client may not wish to be

Table 5. Consumer GIS case study 1 (AT&T Wireless, 2005)

Application Area	Friend Finder
Summary	AT&T Wireless Find People Nearby
Push/Pull Services	Find friends nearby; add and delete people on list; directions to restaurants, bookstores, etc.; locate people, call or SMS friends; hide your location from friends
Geographic Coverage	North America
Positioning	Network based (location of cellular transmission tower most recently contacted)
Mobile Devices	GSM/GPRS mobile phone
Platform	-
Communications Network	Need a phone with an mMode plan and within AT&T Wireless GSM/GPRS network coverage
Comments	Privacy is a major issue

Table 6. Consumer GIS case study 2 (PARAMOUNT, n.d.; Lohnert, Wittmann, Pielmeier, & Sayda, 2001; Lohnert, Mundle, Wittmann, & Heinrichs, 2004)

Application Area	Tour Guide
Summary	PARAMOUNT Public Safety & Commercial Info-Mobility Applications & Services in the Mountains. The purpose of PARAMOUNT is to improve information access, navigation, and safety for mountaineers/hikers. It will help mountaineers to locate, orient, and navigate in unfamiliar terrain, and improve safety and emergency response services.
Push/Pull Services	INFOTOUR delivers information on points of interest, navigation; access to weather forecasts, etc. SAFETOUR delivers safety information on weather, fire warnings, etc.; provides avalanche forecasting; tracks the user's location, sends emergency calls, routes rescue teams to aid mountaineers DATATOUR tracks willing users, updates trail network information, captures information on points of interest or trail difficulty, etc.
Geographic Coverage	Mountainous terrain, outdoors; Pyrenees and Alps
Positioning	GPS
Mobile Devices	Smartphones and PDAs
Platform	-
Communications network	GSM/GPRS, UMTS (in the future)
Comments	Problem of network coverage in mountainous regions; problem of GPS availability in rugged terrain

contacted every time he or she walks past Starbucks or McDonalds). The information delivered to the client may be in the form of text, audio, maps/graphics, or multimedia information. Before the implementation of LBSs, there was no support for the mobile user, except in printed form.

LBSs maintain three common data types: (i) base data on a region (e.g., street network, aerial photographs); (ii) point of interest (POI)

data (e.g., museums, cinemas, restaurants, ATMs, etc.); and (iii) the location of the mobile user. They also involve a much longer value chain, including data providers, network providers, GIS software vendors, application developers, and so forth (McKee, 2004, p. 152). Such services rely on the delivery of location-based information, according to user profiles, preferences, and characteristics. As they continuously record the location of the mobile user

and already have their personal details, privacy and security are a major consideration for location-based services—they are, after all, the closest manifestation of ‘big brother is watching’ concerns. Many services address the privacy issue by allowing the user to opt-in or opt-out of different services and levels of service at any time. However, LBS applications store vast amounts of personal information, and there is need to provide security (encryption) during transmission. The user must also be assured that service providers will not use this information in any other way or pass it onto a third party (the EU has specific legislation covering this; see also Vodafone UK Privacy Policy).

CHALLENGES

The value and benefits of MGISs are apparent from the applications listed above. Yet despite the advantages and the general availability of the technology by the year 2000, they are not yet widely deployed. Why? A range of technical and non-technical issues are involved. These include: integrating MGIS architecture, hardware issues with mobile devices, issues with positioning technologies, issues related to the wireless communications network, and mapping issues.

Integrating MGIS Architecture

MGIS applications depend on a relatively complex set of components. The first challenge is to ensure that the various components that make up such applications can communicate. This is a significant problem given that different mobile devices, networks, GIS software, data servers, and so forth utilise different interfaces and protocols (Chen, 2004). The second challenge is to ensure that MGISs can be integrated with existing (non-mobile) systems (e.g., billing systems) (McKee, 2004; Spinney, 2003)—hence,

the need for standard APIs and protocols (e.g., XML, GML, Java, OSA/Parlay API, CORBA). Such standards ensure that each MGIS application can access distributed and heterogeneous data sources, distributed tools, and services, regardless of the underlying platform, hardware, vendor, network, positioning technology, data formats, and so forth (Zadorozhny & Chrysanthis, 2004, p. 147). Standards also allow for MGISs to be easily implemented with minimal cost and effort (McKee, 2004). MGIS interoperability in its various forms is an *essential* requirement for success—it is not optional.

Hardware Issues with Mobile Devices

Mobile devices are typically regarded as underperforming and overpriced compared to a desktop PC or even a laptop. While the small screen (including its poor resolution and limited colour palette) and limited memory are often cited as the chief failings of these devices, experience also suggests that poor battery life is a major problem—a device with only three hours of continuous use capability is not likely to endear itself to consumers (especially when the battery is not easily replaceable by the user). Some solutions, however, are beginning to emerge (e.g., PDA Trip Project, 2005). Added to these limitations is the fact that these screens can be easily damaged and the devices are easily lost.

The small screen size means that special consideration must be given to the interface design of MGISs. Buttons, tabs, menus, lists, maps, text, images, and so on must generally fit into a much smaller area. Standard Web pages and Windows applications are inappropriate for these devices. The information on such devices must be readable under all conditions (day and night, indoors and outdoors) and so requires enhanced colour contrast.

Table 7. Characteristics of different types of mobile devices³

	Resolution	Colour	Processor	Storage	Memory
WAP phone	96 x 32, 24 x 84, 60 x 80 – 120 x 160, 132 x 176, 176 x 220	Monochrome, greyscale; 256 colour; 4096 colour; 65k 16 bit TFT LCD screens	?	? Limited	Limited; 1, 2, 3, 5, 7 MB
Smart phone	Up to 208 x 320	3375-65536 colours; 64k; 16 bit 65k TFT reflective display	33, 130, 144 MHz	Greater storage; 32; 64 MB Flash ROM; 2GB memory stick	1.7, 4, 5, 8, 10, 16, 32, 64, 128 MB
PDA	160 x 160, 240 x 320, 320 x 320, 320 x 480, VGA resolution 480 x 640	Monochrome; 65k 16 bit colour; 65k colour transfective TFT screens	126, 200, 300, 400, 624 MHz	7.2, 10, 20, 32, 48, 128, 256, 512 MB, 1GB, 4GB	8, 25, 32, 64, 92, 128, 152 MB

Nor do small-screen mobile devices support the same data input methods as large-screen devices such as laptops and desktop computers. While some PDAs have QWERTY keyboards, many do not, relying instead on hand-writing recognition or digital on-screen keyboards for entering text. They have touch screens and utilise pointing devices instead of a mouse. Some also support voice entry (Peng & Tsou, 2003, p.456). Smartphones also utilise touch screens, pointing devices, and on-screen keyboards, as well as having a phone keypad for text entry. WAP phones generally do not have touch screens, on-screen keyboards, or utilise pointing devices. They have much more restricted data entry capabilities. Hence, different devices require different interface designs and support different methods of user interaction.

The fact that there are many different devices—each with different characteristics and capabilities, and supporting different operating systems, mark-up languages, and protocols—is another significant issue (see Table 7; Dao, Rizos, & Wang, 2002). These variations impact negatively on the development of the industry, as they slow down application development,

and multiple platform-specific versions of software prevent scale economies from being realised.

Mobile GIS must be able to support a wide range of devices with different capabilities. It should also be noted that despite the limitations of current devices, there are many high-end mobile phones and PDAs that can adequately support such applications. Hence, extensive ‘proof of concept’ type hardware development is not required. Furthermore, the capabilities of such devices are improving rapidly, although (as it is the defining feature of the device type) small screen size will continue to be an issue. This requires that applications must be designed in such a way that if a device cannot support all the features of an application, the application must fail gracefully (e.g., if a device can be locationally aware, but cannot display maps, it still must operate on text only and not crash).

Issues with Positioning Technologies

A wide variety of positioning technologies are available for MGIS development, ranging from

Table 8. Positioning technologies and their limitations (Readman & Mojarrahi, 2004; Tanner, 2002)

Positioning Technology	Accuracy	Limitations
Cell ID	50m-20km	Greatest in rural areas; often around 200m in urban areas
Cell Triangulation, e.g., EOTD	50-500m	Within a cell sector; often 100m
GPS	10m	Not for indoor use; problem of urban canyons
AGPS	5m	For high-end users
DGPS	<1m	Not for indoor use; problem of urban canyons
Galileo (2008-)	1m	Not for indoor use; problem of urban canyons
Bluetooth	3m	Indoor only, short range

cell-based positioning (e.g., Cell ID, AOA, TDOA, EOTD) and satellite positioning (e.g., GPS), to wireless networks and radio frequency ID (RFID). Hence, applications must support a variety of positioning technologies. The choice is mainly cost driven. Most field-work applications rely on GPS-based positioning, as positional accuracy is important and the cost of the device is not usually a deterrent. In these applications the mobile device establishes its own position. In contrast, for LBS the position is usually fixed by the network, as consumers are generally not expected to be willing to pay the higher price for a mobile device which would be capable of fixing its own position.

Each positioning technique has its own limitations (see Table 8). For example, while the use of GPS is a much more accurate technique than cell-based positioning, they have limitations in indoor and urban environments. However, as GPS receivers become more compact and affordable, they are more likely to be incorporated into a wider range of mobile devices.

The positional accuracy of these different techniques is also quite varied, with some methods such as Cell ID being coarse to the point where the location of a user may be quite misleading (e.g., a map showing the location of the user may place the user in the wrong street or in a building when they are actually driving on the road). Positional accuracy is also influ-

enced by map scale, being quite significant on a 1:10,000 map (Gartner & Uhrlirz, 2001; Uhrlirz, 2001). The widely used method of symbolising the location of the user (e.g., the use of cross hairs) also implies a false accuracy (Gartner & Uhrlirz, 2001; Pospischil, Umlauf, & Michlmayr, 2002, p. 146; Uhrlirz, 2001).

Different applications have different accuracy requirements, and there are applications for which coarse positional accuracy is adequate (see Table 9). Supplementary services can sometimes be used to improve locational accuracy, as is often done within car navigation kits, using the odometer and steering angle if the vehicle goes into an urban canyon. This is used to estimate the position until another fix is found—some heuristic logic would be involved as well, in that the car will only drive on roads.

As many of these positioning techniques are based on cellular phone networks, they cannot be used in areas where coverage is not available. Hence, GPS-based positioning is regarded as superior because of its inherent greater

Table 9. Positional accuracy requirements⁴

Applications	Required Accuracy (m)
Concierge services	100
Outdoor navigation	25
Indoor navigation	<10
Tour guide	30
Emergency services	100
Incident investigation	<10
Facility maintenance	<10

accuracy and better availability in rural areas (Mountain & Raper, 2002, pp. 5-6). In order to overcome this problem of the coverage of mobile networks and to improve the accuracy of cell-based techniques (as well as to combat the limitations of GPS), many applications utilise hybrid positioning techniques (e.g., Cell ID and GPS) (Mountain & Raper, 2002).

Issues Related to the Wireless/ Communications Network

Existing wireless environments for MGIS tend to have the following characteristics (Peng & Tsou, 2003, p. 467):

- **Less Bandwidth:** Bandwidth is the amount of content that can be passed down the narrowest point of a network at a given time.
- **More Latency:** Latency refers to the length of time it takes the content to flow a certain distance. It is a measure of the amount of time it takes for a request to make it from the client and back (“ping”).
- **Less Connection Stability:** The mobile user in motion has to handoff from cell to cell, which could cause information drops during the handoff process. While this should be handled by the connection protocols, it means that the information has to be resent a large number of times—effectively reducing available bandwidth.
- **Less Predictable Availability.**

The limited bandwidth of wireless networks, compared to wired networks, is a major issue for MGISs. Given the large volumes of geographic information and the need for real-time response, MGISs require fast transmission and rendering of spatial information. 2G networks such as GSM typically have a bandwidth of 10kbps and 2.5G networks (e.g., GPRS) up to

115kbps. 3G networks currently being implemented offer vastly improved rates of up to 2Mbps (Peng & Tsou, 2003, pp. 458-467). 2.5G and 3G networks also have the advantages of streaming, being always on, and supporting the transmission of multimedia information (Gartner & Uhlig, 2001; Nissen, Hvas, Münster-Swendsen, & Brodersen, 2003; Uhlig, 2001). Caching of surrounding data while the user is looking at current data, or when they are in range of a larger capacity network (such as when a worker goes to a Wi-Fi hotspot and caches a lot of data and then updates and refreshes the data using GPRS) goes some way to addressing this problem.

Coverage is another major issue with cellular networks, as without coverage, no data or network-based positioning information can be received or transmitted. In this case, the user must rely on the information stored locally in the mobile device. Given the limited memory and storage of many mobile devices, this has been an issue. Technological improvements, however, are significant (e.g., availability of 4gb+ memory cards). Lack of network coverage also reduces the real-time data updating benefit of field applications of MGIS. The client-server model used has important implications here, and the most suitable model depends on the types of mobile devices, network characteristics, and required functionality (Mountain & Raper, 2002, pp. 5-7).

Mapping Issues

The change from 17” to 3” screens and the poor bandwidth in wireless communications create major problems for GIS developers. While acknowledging that there are several significant mapping issues (discussed below), it should not be assumed that traditional GISs and MGISs always have the same cartographic objectives or requirements; clearly, the appli-

cation requirements are very different—simple maps or even text-only output are quite acceptable for many MGIS queries (e.g., routing instructions might look ‘nicer’ in a colour map, but plain text on a monochrome screen is quite acceptable). With the continuing advances in ‘text to speech’ applications, audio may be more practical in some circumstances as the driver is not required to take his/her eyes off the road.

Maps form a major component of MGISs and must not only fit within the small screen, but also within the application interface, with its menus, buttons, and so forth. Maps in Mobile GISs need to be simple, uncluttered, and easy to read at arm’s length (Hjelm, 2002, p. 247). In order to combat the bandwidth problem, many applications have used predefined, static, scanned raster maps (Pospischil et al., 2002). However, while scanned raster maps result in lower storage, transmission, and rendering costs, they allow for little manipulation or interactivity (Abowd et al., 1997; Hjelm, 2002, p. 257; Reichenbacher 2001a, 2001b, 2003, p. 1317). Rescaling also results in information loss unless data pyramids are used—adding to the storage overhead. Raster data layers also require high storage and transmission costs.

With the improvements offered by 2.5G and 3G networks, there has been an increasing focus on the use of vector data, which allow for lower storage and transmission costs, easier manipulation of data, a wider range of applications, links to a database, and adaptive, dynamic, and interactive mapping (Abowd et al., 1997; Long, Kooper, Abowd, & Atkeson 1996, p. 104; Reichenbacher, 2001a, 2001b). Traditional vector data formats, however, require much greater processing for map display and are not suited to mobile mapping; new data types and formats are required (De Vita, Piras, & Sanna, 2003; Hjelm, 2002, p. 257; Reichenbacher, 2001a, 2001b).

Adaptive mapping refers to the situation where maps are created on the fly according to the user’s context, preferences, and characteristics (age, language, culture, preferences, etc.). Traditionally, context is regarded as the user’s location, but other attributes such as weather, time, bandwidth, and mobile device can be included (Reichenbacher, 2001a, 2001b; Zipf, 2002). It is the context and user profiles that determine the content, level of detail, and degree of generalisation of the maps in adaptive mapping (Reichenbacher, 2003).

Other important cartographic issues include the need for common rules and standards for mobile cartography. Symbology needs to be automatically assigned according to widely understood colours and symbols, taking into consideration cultural conventions and the need to maintain high contrast (Nissen et al., 2003; Zipf 2002). Labels, varying according to the map scale and level of detail, must also be provided in the preferred language of the user. The small screen taxes any automated name placement routine for clutter and overlap (Reichenbacher, 2003).

TOWARDS SOLUTIONS

Open standards are required for interoperability to ensure that MGISs can operate across different wireless networks, platforms, devices, positioning technologies, data types, and so forth. They also allow for the various components of an application to communicate and to be easily integrated with existing systems. There are a number of standards bodies working on standards for MGISs, the most significant of which are the Organisation for Standards (ISO), Open Mobile Alliance (OMA), and Open GIS Consortium (OGC). For example, the Location Interoperability Forum (LIF), which is now part of the OMA, developed the mobile location

protocol (MLP). The MLP is an open interface for determining the position of a mobile device, regardless of positioning method and network (i.e., it interfaces the location server and application server) (Zadorozhny & Chrysanthis, 2004, pp. 151-153). The OGC OpenLS GeoMobility Server (GMS) defines open interfaces for core spatial services (McKee, 2004, pp. 160-168; OpenLS, 2004).

In order to address the issue of small screen size, some applications only support particular devices or devices above a certain resolution. For example, the GiMoDig project only supports devices with a resolution of 180 x 180 or above, Java ME capability, and suitable memory (Nissen et al., 2003). It is also recommended that applications should reduce the length of their pages and utilise hypertext links. Shorter pages and a deeper hierarchy are recommended for the wireless Web (Kacin, n.d.). Using a hierarchy of maps, interactive mapping, and multimedia information also allows for optimization of content for small screens (Gartner & Uhlig, 2001; Pospischil et al., 2002; Uhlig, 2001). Maps must also be simple, uncluttered, and have minimal detail. Hence, many maps are predefined and static (Gartner & Uhlig, 2001). While screen size will continue to be a problem, technological improvements in mobile devices will result in better resolution, colour range, battery life, and so forth. Owing to the difficulty of text entry on many mobile devices, voice entry remains an important, but as yet underdeveloped, means of input.

Improvements in positioning technologies will result in greater positional accuracy, as will the use of more precise positioning techniques (e.g., Galileo). Accuracy can further be improved through the use of hybrid positioning techniques, user input, the use of passive landmarks and active sensors (e.g., Bluetooth), as well as video and augmented reality (Gartner, 2004; Gartner & Uhlig, 2001; Uhlig, 2001).

Furthermore, it is important to acknowledge the uncertainty of the user's location—the symbology utilised should indicate the accuracy of the location or error term (e.g., using a circle of varying size to indicate the error term). In addition, rather than showing a precise location, it is possible to indicate the region or street within which the user is located (Gartner & Uhlig, 2001; Pospischil et al., 2002; Uhlig, 2001). The symbology used should also indicate when real-time positional information is not available.

The communications bandwidth problem is partially solved by the use of 2.5G and particularly 3G technology. They also allow for a larger range of client-server models (Mountain & Raper, 2002, p. 4). Adaptive mapping and generalisation have reduced the bandwidth problem further, by reducing the volume of information transmitted (Gartner & Uhlig, 2001). The improved caching ability of current devices also reduces network traffic and the problem of network coverage.

Furthermore, the client-server model chosen has an impact on the network traffic. Server-side applications place a greater load on the network, and require high transfer rates and streaming mode. However, they also support a wider range of devices, particularly lower end devices, and have greater functionality, with support for interactive and dynamic mapping. Client-side applications, while reducing network traffic and allowing offline processing, require high-end devices; the software on these devices has limited functionality and they favour prepared, static maps (Mountain & Raper, 2002, p. 6; Reichenbacher, 2001b; Gartner & Uhlig, 2001). It is relatively simple to have a raster background with a vector layer overlaid so that most of the data is static, but the 'relative' information is in vector format, so that all the vector advantages are available.

Vector data formats offer greater potential to MGISs than raster data formats, and so a number of vector data formats have been developed especially for mobile devices; examples include Compact GML (cGML), SVG Basic (for PDAs), and SVG Tiny (for mobile phones) (De Vita et al., 2003; Reichenbacher, 2003). Currently, most LBSs utilise data in GML and then convert it to SVG for display (e.g., PARAMOUNT; GiMoDig).

The mobile mapping challenges have also led to the development of new forms of representation such as focus maps, bird's eye perspectives, floor plans, and the use of multimedia information (Gartner, 2004). Focus maps reduce transmission costs by providing a detailed representation of features within a certain distance of the user and a coarse level of detail for more distant features (Zipf & Richter, 2002). Adaptive mapping also reduces transmission and rendering overheads.

With regard to label placement, one way to mitigate the problem of cluttered and overlapping labels is to provide few details on overview maps and to require the user to zoom into the maps for more information (Heidmann, Hermann, & Peissner, 2003). This issue can also be addressed through the use of labelling algorithms, but is most frequently solved through the use of tool tips and hot spots. This greatly reduces the number of labels required, and the user can still obtain information on features by passing the cursor over the feature or clicking on it (Gartner & Uhrlirz, 2001; Uhrlirz, 2001).

FUTURE DIRECTIONS

With improvements in technology (software and hardware), many of the challenges mentioned above will be reduced or eliminated in the near future. For example, with the development of mobile devices with greater memory, better battery life (including user-replaceable

batteries), more powerful processors, and screens with higher resolution, many of problems of mobile devices will no longer be applicable. While devices are constantly improving, however, small screen size will remain an issue. A reduction in the cost of such devices will also allow for greater adoption of this technology and better user acceptance.

The implementation of 3G and better networks will also alleviate bandwidth problems and support a greater range of client-server models. The use of more precise positioning technologies (e.g., DGPS, Galileo), and improvements in the availability and costs of such technologies will also improve the locational accuracy of applications. Most significant here will be the increasing availability of mobile devices with built-in GPS receivers.

These technology improvements will make Mobile GISs of even greater importance to enterprises in achieving operational efficiencies in field operations generally and with significant gains in 'field force automation' in particular. However, while technology improvements are a prerequisite, more widespread consumer adoption of this technology will be hindered unless applications of greater utility to the consumer (as well as more "compelling content") become available.

REFERENCES

- Abowd, G., Atkeson, C., Hong, J., Long, S., Kooper, R., & Pinkerton, M. (1997). Cyberguide: A mobile context-aware tour guide. *Wireless Networks*, 3, 421-433.
- AT&T Wireless. (2005). *Find people nearby*. Retrieved March 29, 2005, from <http://www.attwireless.com/personal/features/organisation/findfriends.jhtml>

- Chen, A. (2004). *Open standards will evolve location-based services*. Retrieved November 30, 2004, from http://www.eweek.com/print_article2/0,2533,a=131075,00.asp
- Conolly, N. (2001). Software for mobile mapping. *GIS User* 47. Retrieved March 29, 2005, from http://www.gisuser.com.au/GU/content/2001/GU47/gu47_frame.html
- Dao, D., Rizos, C., & Wang, J. (2002). Location-based services: Technical and business issues. *GPS Solutions*, 6, 169-178.
- Daratech. (2005). *Leading manufacturers validate PLM though confusion remains*. Retrieved from <http://www.daratech.com/press/releases/2005/050228.html>
- De Vita, E., Piras, A., & Sanna, S. (2003). *Using compact GML to deploy interactive maps in mobile devices*. Retrieved from <http://www20003.org/cdrom/papers/poster/p051/p51-devita.html>
- ESBI Computing. (2005a). *Mobile GIS: Benefits*. Retrieved March 17, 2005, from www.esbic.ie/geobusiness/mobile_gis/GIS_Benefits.htm
- ESBI Computing. (2005b). *Mobile GIS: Application to industry*. Retrieved March 17, 2005, from www.esbic.ie/geobusiness/mobile_gis/GIS_Industries.htm
- ESRI. (2005). *Mobile GIS*. Retrieved March 17, 2005, from www.esri.com/software/arcgis/about/mobile.html
- Gartner, G. (2004). Location-based mobile pedestrian navigation services—The role of multimedia cartography. In *Proceedings of ICA UPIMap 2004*, Tokyo. Retrieved from <http://ubimap.net/upimap2004/html/papers/UPIMap04-B-03-Gartner.pdf>
- Gartner, G., & Uhlirz, S. (2001). Cartographic concepts for realising a location based UMTS service: Vienna city guide “LOL@”. *Proceedings of the 20th International Cartographic Conference* (vol. III, S.3229-3239), Beijing. Retrieved from http://lola.ftw.at/homepage/content/a40material/Vienna_City_Guide_LoLa.pdf
- Geological Survey of Western Australia*. (2005). Retrieved March 15, 2005, from http://www.linkspoint.com/docs/GSWA_CS.pdf
- Giaglis, G., Kourouthanassis, P., & Tsimakos, A. (2002). Towards a classification for mobile location services. In B. Mennecke, & T. Strader (Eds.), *Mobile commerce: Technology, theory, and applications* (pp. 64-81). Hershey, PA: Idea Group Publishing.
- Hassin, B. (2003). *Mobile GIS: How to get there from here*. Retrieved March 29, 2005, from <http://gis.esri.com/library/userconf/proc03/p0988.pdf>
- Heidmann, F., Hermann, F., & Peissner, M. (2003). Interactive maps on mobile, location-based systems: Design solutions and usability testing. In *Proceedings of the 21st International Cartographic Conference (ICC)* (pp. 1299-1305), Durban, South Africa.
- Hjelm, J. (2002). *Creating location services for the wireless Web: Professional developer's guide*. New York: John Wiley & Sons.
- Kacin, M. (n.d.). *Optimizing Web content for handheld devices*. Retrieved from <http://www.wirelessdevnet.com/channels/pda/features/handheldcontent.html>
- Karimi, H. A., & Hammad, A. (2004). *Telegeoinformatics: Location-based computing and services*. New York: CRC Press.
- Lohnert, E., Mundle, H., Wittmann, E., & Heinrichs, G. (2004). Wireless in the Alps: An LBS prototype for mountain hikers. *GPS World*, 15(3), 30-37.

- Lohnert, E., Wittmann, E., Pielmeier, J., & Sayda, F. (2001, September 11-14). PARAMOUNT—Public Safety & Commercial Info-Mobility Applications & Services in the Mountains. In *Proceedings of the 14th International Technical Meeting of the Satellite Division of the Institute of Navigation (ION GPS 2001)*, Salt Lake City, UT (pp. 319-325).
- Long, S., Kooper, R., Abowd, G., & Atkeson, C. (1996, November). Rapid prototyping of mobile context-aware applications: The Cyberguide case study. In *Proceedings of the 2nd ACM International Conference on Mobile Computing and Networking (MobiCom '96)*, Rye, NY (pp. 97-107). New York: ACM Press.
- Longley, P., Goodchild, M., Maguire, D., & Rhind, D. (2001). *Geographic information systems and science*. New York: John Wiley & Sons.
- McKee, L. (2004). *LBS interoperability through standards*. In J. Schiller & A. Voisard (Eds.), *Location-based services* (pp. 149-171). Amsterdam: Elsevier.
- Mountain, D., & Raper, J. (2002, September 18). Location-based services in remote areas. In *Proceedings of the Association of Geographical Information* (Paper B5.1, pp. 1-9). Retrieved from <http://www.soi.city.ac.uk/~dmm/research/pubs/B05.3.pdf>
- Nissen, F., Hvas, A., Münster-Swendsen, J., & Brodersen, L. (2003). *KMS, National Survey and Cadastre—Denmark: Small-display cartography*. Retrieved November 30, 2004, from http://gimodig.fgi.fi/pub_deliverables/D3_1_1.pdf
- North London Strategic Alliance. (2005). Retrieved from <http://www.publictechnology.net/print.php?sid=861&POSTNUKESID=1a6d6da2c722c833fb65eddf8ff072e3>
- Open LS. (2004). *Open GIS location services*. Retrieved December 15, 2004, from <http://www.opengeospatial.org/specs/?page=specs>
- PARAMOUNT. (n.d.). Retrieved from <http://www.paramount-tours.com>
- PDA Trip Project. (2005). Retrieved March 29, 2005, from <http://www.7volts.com/travel.htm>
- Peng, Z., & Tsou, M. (2003). *Internet GIS: Distributed geographic information services for the Internet and wireless networks*. New York: John Wiley & Sons.
- Pospischil, G., Umlauf, M., & Michlmayr, E. (2002). Designing LOL@, a mobile tourist guide for UMTS. *Lecture Notes in Computer Science*, 2411, 140-154.
- Readman, D., & Mojarabi, B. (2004). *Location based and communication systems*. Unpublished manuscript, University of Queensland, Australia.
- Reichenbacher, T. (2001a). Adaptive concepts for a mobile cartography. *Supplement Journal of Geographical Sciences*, 11, 43-53.
- Reichenbacher, T. (2001b). The world in your pocket—Towards a mobile cartography. In *Proceedings of the 20th International Cartographic Conference*, Beijing, China. Retrieved from http://citeseer.ist.psu.edu/cache/papers/cs/23234/http:zSzzSzwww.lrz-muenchen.dezSz~t583101zSzWWWzSzpublicationszSzreichenbacherzSzICC2001_Paper.pdf/reichenbacher01world.pdf
- Reichenbacher, T. (2003, August 10-16). Adaptive methods for mobile cartography. In *Proceedings of the 21st International Cartographic Conference (ICC)* (pp. 1311-1322), Durban, South Africa. Retrieved from <http://www.carto.net/geog234/readings/>

reichenbacher_mobile_cartography_durban_2003.pdf

Smyth, C. (2000). *Mobile geographic information services: Turning GIS inside out*. Retrieved from <http://www.giscience.org/GIScience2000/invited/Smyth.pdf>

Spinney, J. (2003). *A brief history of LBS and how OpenLS fits into the new value chain*. Retrieved from http://www.esbic.ie/geobusiness/Mobile_GIS/overview.htm

Tanner, J. (2002). Location-based services: Where it's at. *Wireless Asia*, (November), 22-24.

Uhlerz, S. (2001). Cartographic concepts for UMTS-location based services. *Proceedings of the 3rd Workshop on Mobile Mapping Technology*, Cairo, Egypt. Retrieved from http://lola.ftw.at/homepage/content/a40material/Cartographic_Concepts_for_UMTS_Location_based_Services.pdf

Zadorozhny, V., & Chrysanthis, P. (2004). Location-based computing. In H. Karimi & A. Hammad (Eds.), *Telegeoinformatics: Location-based computing and services* (pp. 145-170). Boca Raton, FL: CRC Press.

Zipf, A. (2002). User-adaptive maps for location-based services (LBS) for tourism. In K.

Woeber, A. Frew, & M. Hitz (Eds.), In *Proceedings of the 9th International Conference for Information and Communication Technologies in Tourism* (ENTER 2002), Innsbruck, Austria. Retrieved from <http://www.eml-development.de/english/homes/zipf/ENTER2002.pdf>

Zipf, A., & Richter, K. F. (2002). Using focus maps to ease map reading. developing smart applications for mobile devices. *Artificial Intelligence* (special issue on spatial cognition). Retrieved from <http://www2.geoinform.fh-mainz.de/~zipf/ki-04.2002-zipf.pdf>

ENDNOTES

- ¹ For more information on Mobile GIS software, see Peng & Tsou (2003, pp. 480-492) and Conolly (2001).
- ² Information for this table has been adapted from Peng & Tsou (2003, pp. 493-495). For a broader range of applications, see ESBI Computing (2005b).
- ³ Information for this table was derived from Internet research on currently available mobile phones and PDAs.
- ⁴ Some figures were derived from Gartner (2004).

Chapter LVIII

Mobile Technologies and Tourism

Pramod Sharma

The University of Queensland, Australia

Devon Nugent

The University of Queensland, Australia

ABSTRACT

This chapter examines the potential of mobile technologies for the tourism industry. Mobile technologies have the capacity to address not only the pre- and post- tour requirements of the tourist, but also to support the tourist on the move. It is this phase of the tourist activity upon which mobile technologies can be expected to have the greatest impact. The development of applications for the mobile tourist will allow for the creation of a new range of personalised, location and time specific, value added services that were not previously possible. Before such applications can be widely deployed, however, some fundamental technical and business challenges need to be addressed. Despite these challenges, mobile technologies have the potential to revolutionise the tourist experience, delivering context specific services to tourists on the move.

INTRODUCTION

Tourism is one of the world's largest and most rapidly expanding industries, contributing over 10% to global GDP (WTTC, 2003). Information and communication technology (ICT) has played a critical role in its development, as evidenced for example by the development of massive global distribution systems (GDSs) and airline computerised reservation systems (CRSs), as well as enterprise systems, such as

FIDELIO for the hospitality industry. Until relatively recently most ICT applications dealt with enterprise operations of the B2B type. The Web has had a "liberating" effect by making possible the introduction of B2C applications (e.g., using a Web site for the distribution of product information, destination promotion, online bookings, and e-commerce in general).

The B2C applications have been relatively "static" systems offering support in the pre- and, to some extent, post-tour phase of a trip.

For example, they support information retrieval on places to visit, online reservations, personal Web pages, and so forth. A key gap in ICT applications, however, has been the lack of support for the tourist on the move (e.g., “Where is the nearest hotel?” “How do I get to the museum from here?” etc.). Solutions such as kiosks, Internet cafes, and the increasing availability of Internet connections in hotel rooms address only part of the problem, as they still require the tourist to be wired (“tethered”). The answer lies in developing applications which address the mobility issue *directly* by delivering applications on wireless-enabled platforms such as mobile phones and PDAs to tourists while they are on the move. Herein lies the challenge: at its simplest the issue is one of providing information-rich, “bandwidth guzzling” content (colour photos, graphical and audio-visual content) via “capacity-challenged” hardware platforms and communications infrastructure. Thus, exhortations to develop applications that provide “personalised” services “in context” are not surprising, both because they are desirable (CSTB, 2003) and also because for the foreseeable future, hardware limitations will dictate it.

Despite the challenges, the vision is clear enough:

... the opportunity for providing location dependent information and reservation is critical for tourism[,] and the constantly moving consumer ... will support a whole new way of communicating, accessing information, conducting business, learning and being entertained while on the move...With access to any service anywhere, anytime from one terminal, the old boundaries between communication, information, media and entertainment will gradually disappear, offering convergence between technologies and tourism services. (Buhalis, 2003, p. 323)

In a sense the issue is transitional, for it is not difficult to foresee that, in the longer term, wireless forms of communication will be the dominant technology, and that technology and business challenges will be addressed. However, regardless of the duration of the transitional period, certain fundamental issues need to be addressed: Why are mobile technologies of interest to tourism? What are the key components of a mobile system? What kind of services are required by the mobile tourist? What are the issues involved in applications development for mobile services? Case studies of existing applications in tourism will be used to illustrate some of these points. This chapter will answer these questions and evaluate the potential impact of mobile technologies on tourism.

INFORMATION REQUIREMENTS FOR MOBILE TOURISTS

Using “old” and “new” tourism models (O’Looney, 2004) has highlighted some aspects of information use by tourists. In the “old Tourism Model”, tourists used guidebooks, paper maps, and printed media. They were inundated with cluttered and broad-based advertising, and lacked the ability to focus or follow-up on a topic. Furthermore, much of this information was out of date soon after printing. Guidebooks can also be hard to follow: it is not easy to link locations on a map with information in guidebooks, one must read through lots of information in order to get what one wants, and such guides are designed for the “general” tourist (Brown & Chalmers, 2003). Web information searches, when available, essentially mimicked the traditional library service, providing vast amounts of unstructured data. In contrast, the ‘new tourism model’ uses mobile phones, PDAs, laptops, and electronic maps; it provides a personalised, information-rich, and location-

specific experience; multiple media, content-linked advertising; and the new ability to focus or follow-up on a topic. The user can decide how much information they want. However, while this vision is accurate in terms of technology, it is far too “rosy in terms of user acceptance; *much* remains to be done both in terms of enabling technology as well as tourism information content to convert the “new” tourism model into reality.¹

What are the limitations of current tourist information systems? A recent paper (Watson, Akselsen, Monod, & Pitt, 2004, pp. 315-316) identifies three fundamental problems:

1. tourists are overwhelmed by the variety and the volume of information on the Web, and experience information overload;
2. there is little ICT support during the touring phase; and
3. experiences gained during a trip are not easily shared, and reminiscing is rarely supported.

While all of these matters are of research interest, our focus is mainly on issues related to ICT support during the touring phase.

The tourist activity is frequently envisaged as having three stages or phases—each has different information requirements. The first of these is the pre-tour or planning phase, during which the tourist is actively researching products and destinations. The Web, despite its shortcomings, is increasingly becoming the preferred source for information (Epstein, Garcia, & Fiore, 2003, p. 29) on which the decision to travel or not to travel is made. The proliferation of travel and tourism Web sites has meant that a huge volume of variable quality information is available on almost any tourism topic—ranging from destination sightseeing information to information on types of plugs for electrical appliances.² If this frequently results in “information

overload” for the consumer in a typical Internet environment (a desktop PC with a dialup modem, if not broadband, connection), it is not hard to imagine the “system overload” it will lead to in a hardware- and infrastructure-challenged mobile environment.

The third stage is post-tour or reminiscing when there is little need to research information, but there is a need to record and share the tourist’s own information and experiences with family and friends. This reminiscing or recording and sharing of information also has other uses—it is invaluable information for other tourists. Thus it should be noted that the Web permits ‘word of mouth’ dissemination through the various “user review” or “feedback” facilities; available evidence suggests that tourists rate the experiences of other tourists as equal to, if not better than, those of ‘experts’ (Brown & Chalmers, 2003). An additional aspect of interest to applications developers is the use of this information as input into travel recommender systems (TRSs).

It is our view that it is in the second stage—the touring itself (i.e., when the tourist is on the move)—that ICT in general and mobile technologies in particular can make a significant contribution to the tourism industry by meeting the information needs of the tourist. These information needs take on a certain urgency when we realise how stressful lack of information can be when the tourist may be in unfamiliar (“threatening”) surroundings and also when plans may need changing in light of changing circumstances. The need for flexibility is an important aspect of the tourist experience, as tourists enjoy being able to change their plans—over planning can produce unhappy experiences (Brown & Chalmers, 2003). A number of promising applications, such as LOL@, Webpark, and The Electronic Guidebook, address all three phases of tourist activity (see Tables 3 and 4).

What kinds of information does the tourist need during the touring phase? Two broad categories can be envisaged. The first category consists of ‘administrative’ information regarding the tour itself. These include confirmation of or changes to travel arrangements and bookings (e.g., delayed or cancelled flights, changes to accommodation arrangements, or other changes to itineraries). While e-mails are clearly superior for their information content, SMS has been embraced more widely, as such service can be sent to even the very basic handsets, which generally cannot handle e-mails. This service is already being used by the various airlines. Mobile technologies have clearly represented a great advance in keeping the tourist updated—industry developments such as e-ticketing reinforce this further.

The second, a broader category, usually containing a large volume of information, is made up of destination information. Traditionally the content has been the justification for the creation and maintenance of Destination Management Systems (DMSs) managed by both semi-governmental, as well as private tourism authorities (e.g., national, state, regional, or district tourism boards). Destination information for the mobile tourist is made up of three categories (Eriksson, 2002; for a very ambitious proposal involving a richer content, see O’Looney, 2004):

- information about tourism products (about restaurants, hotels, museums, amusement parks), and various directory (“yellow pages”) services;
- information about transport networks and traffic; about transport network infrastructure (e.g., roads, railways and airports, timetables, etc.); and
- information about locations and positions (to be able to provide answers to: Where is the nearest x? How do I get there?),

information on points of interest (POIs), navigation/route guidance, tracking friends/other members of the party.

To be of value, this information has to be current, context sensitive, personalised, and preferably in the chosen language. Currently, GIS service and content providers can meet most of these requirements on fixed/wired/non-mobile systems; however, it is a big task to deliver these services on mobile systems. The load on the system also varies with the behaviour of the tourist. Thus a tourist who has undertaken all research and downloads off a fixed Internet connection prior to starting the tour and merely accesses the Internet during the tour for a “top-up” or updates, imposes a significantly lower load than the tourist who starts the process during the tour. In the latter case, a series of factors (the learning process, longer connect times, more information downloads, and the limitations of mobile hardware and the communications infrastructure) combine to make it a potentially difficult experience.

Are mobile tourism information systems/services different from fixed Internet applications? At first glance it is reasonable to assume the mobile tourism services are merely the “light” versions of standard Internet tourism applications and should pose no particularly difficult or unique problems. Unfortunately this is not the case, for both the communications network (wired *vs.* wireless systems), as well as client terminals (large screen, mains powered *vs.* small screen battery powered), pose major problems for the development of mobile applications for tourists. While the problems relating to the communications network may be eliminated with the advent of universal 3G or better phone systems, the hardware limitations of mobile devices (especially the small screen constraint) are not likely to be eliminated in the foreseeable future. Furthermore, mobile tour-

ism information systems allow new kinds of services and fulfil user requirements that are not met via fixed networks.

EMERGING SOLUTIONS FOR THE MOBILE TOURIST

Despite the challenges, significant applications do exist either as prototypes or as operational systems. In order to provide a context for an examination and evaluation of these applications, we need to address some issues: What are the ways of getting information and services to the tourist? What is the basic architecture of these systems?

How do we get the information and services to the mobile tourist? Two situations can be envisaged (see Table 1):

1. **The Unconnected Mobile Tourist:** This is the situation where the tourist is not connected to a communications network and is using either a “dumb” or a “smart” electronic guidebook type of product—dumb or smart refers to whether the mobile device is context aware in terms of

location (i.e., is able to determine its own geographic location). Dumb devices are units which allow access to “softcopy” versions of travel guides, books, and so forth; smart devices are the various GPS-enabled PDAs (e.g., NavMan units, Garmin IQUE3600) or GPS-equipped phones.

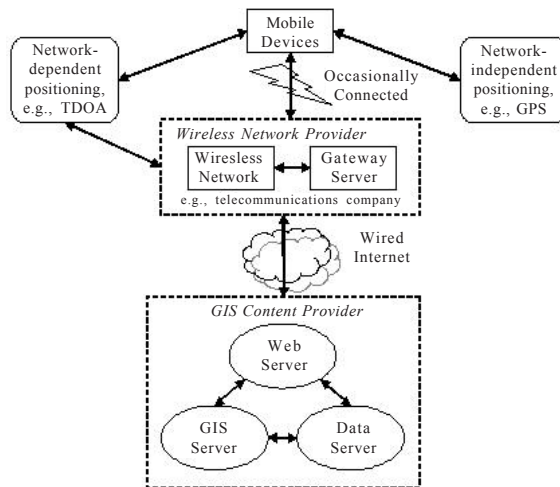
2. **The Connected Mobile Tourist:** This is when the device is connected to a network either by wireless (e.g., Wi-Fi networks (Wireless Internet Network 802.11b) to surf the Internet, check e-mail, and fix position (Epstein et al., 2003)) or to a mobile phone system—the connection to a communications network permits data transfer in both directions. The device may be smart (i.e., it can establish its own position), but even if it is dumb, it may be possible to “fix” its position by the communications network.

While it is the connected mobile tourist who is the focus of this chapter, a brief comment is necessary to illustrate that ICT support of other types (e.g., electronic guides) are possible. As the unconnected dumb mobile electronic guides

Table 1. Types of mobile tourist guides

Type of Mobile Tourist	Location Aware	Description
Unconnected to mobile network	Dumb	No positional information; no data updates (e.g., e-book, digital guidebook)
	Smart	Handset-based positioning (GPS); location-specific information delivery; no data updates (e.g., NavMan)
Connected to mobile network	Dumb	Network positioning; real-time, location-specific information delivery (e.g., Sony Ericsson 910)
	Smart	Handset-based positioning; real-time, location-specific information delivery (e.g., Wherifone)

Figure 1. Components of mobile tourism information systems



have the limitations of mobile devices and do not allow for information updates, they offer few advantages over printed guides and maps (i.e., users have to deal with all the difficulties of using small-screen devices without the full benefits of mobility). In contrast, unconnected smart mobile guides offer the advantage of locating the user in real time and are able to deliver information relevant to that location (although the information is only as current as the last date at which the data on the mobile device was updated).

What is the basic system architecture of a mobile tourism application? Mobile tourism applications are developed around the following components:

- A mobile device (the client terminal), which is capable of being detected by some position-determining technology. The mobile devices tend to be mobile phones (especially those with better, often colour screens), smartphones, and wireless capable PDAs.

- Position-determining technology—the key distinction here is whether the position is determined by the mobile device (e.g., using GPS) or by the communications network (by using one of several triangulation methods).
- GIS content provider (Web server, GIS server, data server).
- Wireless network provider—these are required to communicate with the mobile device.

While superficially similar to standard Internet tourist applications, closer inspection reveals that wireless connected mobile tourist applications are radically different. These applications have the ability to extend the traditional user experience, providing real-time, location-specific, value-added, personalised services that were not previously possible. They are able to support all three phases of the tourist activity, especially the tourist on the move. While touring, the user can access a wide range of additional material and can interact with this information (Semper & Spasojevic, 2002); they can interact with dynamic maps, multimedia content, and text; explore information on a point of interest (POI); record information on places visited (in a travel diary); ask questions about what features are nearby; access the recommendations of other tourists; get directions; and so forth. Most importantly with access to dynamic and up-to-date information, they can ask new kinds of questions such as, “Where is the nearest *open* museum?” (Cheverst, Davies, Mitchell, Friday, & Efstratiou, 2000, p. 18; Watson et al., 2004, p. 317). They combine the advantages of maps and guidebooks with those of dynamic and interactive digital content, which is accessible “anywhere, anytime”. Tourism applications of this kind not only provide a new means of delivering information available on the Web, in guides, and so forth, but also

provide new “innovative” services (Eriksson, 2002, p. 17).

Such applications utilise three common data types: (1) base data (street network, satellite imagery, etc.); (2) point of interest (POI) data (e.g., restaurants, ATMs, cinemas, museums, etc.); and (3) the location of the user. While maps are the most common form of output, data is delivered to the user in a variety of forms—voice, text, images, video, maps, and so on—according to the user’s context (location, device, network, the weather, etc.) and profile (age, interests, language, disability, etc.) (Reichenbacher, 2001a, 2001b, 2003). For example, for navigation when driving, it is more appropriate to deliver audio directions, as the user may not be able to read text or a map while driving. In addition, the user can access a wide variety of textual and multimedia information relating to POIs. The depth to which data is explored is essentially an exchange between content availability and the user profile.

These mobile tour guides can offer a wide variety of services to business and leisure travellers (see Table 2).

Mobile tourists have different requirements from those of tourists searching the Internet. They are often very busy/distracted—looking around, taking in the scenery, navigating, and so forth. They want maps that are easy to read on the small screen, at arm’s length, with easily legible text and high colour contrast for outdoor

use (Hjelm, 2002, p. 247). Nor do they have time to sift through pages of text. Mobile tourists want short, directed text, and the ability to find out further information if they wish; simple interfaces and personalised information delivery; applications that work the first time (‘out of the box’) and every time; and they must offer significant advantages over traditional printed tour guides and Internet applications (Umlauf, Pospischil, Niklfeld, & Michlmayr, 2003). For excellent ‘proof of concept’ examples, see Tables 3 and 4.

Other examples of mobile tourist guides include:

- PARAMOUNT, which was developed to improve information access, navigation, and safety for mountaineers in the Pyrenees and Alps (Lohnert, Mundle, Wittmann, & Heinrichs, 2004; Lohnert, Wittmann, Pielmeier, & Sayda, 2001);
- LOVEUS, which aimed to “provide European citizens with ubiquitous services for personalised, tourism-oriented multimedia information related to the location and orientation within cultural sites or urban settings” (Karagiozidis, Zacharopoulos, Xenakis, Demiris, & Ioannidis, 2003);
- TellMaris, which developed TellMarisOnBoard and TellMarisGuide for leisure boat tourists of the Baltic Sea Region; the main purpose was to explore

Table 2. Some services offered by mobile tour guides

Service	Example
Routing	The best route for visiting the British Museum, National Gallery, and Tate Gallery in a single day.
Navigation	Driving directions from the Marriot Hotel in Brisbane to Toowoomba.
Tour Guide	A pre-defined, one-day tour around Paris.
Concierge Services	Where is the nearest Italian restaurant?
Travel Diary	Record a photo of the Eiffel Tower and comment on the visit.
Travel Recommendations	Access other travellers’ opinions on a particular cycling tour of the Swiss Alps.
Emergency Response	Call for assistance when your hired car breaks down.
Marketing	Receive information on special tour offers.
Tracking	Tracking the location of a tourist along a walking track in a national park.

Table 3. LOL@ mobile tourist guide (Umlauft et al., 2003; Pospischil, Umlauft, & Michlmayr, 2002; Annegg, Kunczier, Michlmayr, Pospischil, & Umlauft, 2002)

APPLICATION AREA: Mobile Tourist Guide	
B2B/B2C: B2C	IMPLEMENTED: On laptop
SUMMARY: LOL@ is a prototype of a mobile electronic interactive tour guide for the city of Vienna. The aim of this project was to develop a location-based multimedia service for UMTS. It supports three kinds of services: pre-tour planning (maps, text, and multimedia information on tours and POIs, filtered according to the user's profile/behaviour); support for the tourist on the move; access to a tour diary (stored on the server) via the Web or PC—the user can record text and multimedia information while on tour, friends can also follow the user's activity.	
PUSH/PULL SERVICES PROVIDED: Push—predefined or user-defined tours; records visited POIs, route segments travelled, and so forth; information (text and multimedia data) on POIs/landmarks; route guidance and navigation; adding to and accessing the electronic tour diary. Pull—records information on visited POIs in tour diary, user location, and orientation.	
GEOGRAPHIC COVERAGE/ENVIRONMENT: Outdoor urban environment, Vienna's First District	
POSITIONING TECHNIQUES: GPS; A-GPS; Cell ID; radio signal propagation; user input; small active or passive senders (e.g., Bluetooth) in the future	
MOBILE DEVICES SUPPORTED: Smart phones, PDAs, and laptops; demonstration version uses a laptop, as PDAs do not provide for full functionality of the map viewer applet or speech recognition software (Pospischil et al., 2002, p. 152); for PDA-like phones 120 x 320 pixels, with Java ME, colour LCD display, and pen input (Annegg et al., 2002); platform comparable to Windows Pocket PC 2002 with Internet browser, 32 or 64 MB memory, virtual keyboard, text recognition, and simple GUI (Umlauft et al., 2003)	
COMMUNICATIONS INFRASTRUCTURE: UMTS or GPRS (always on)	
DESIGN INTERFACE: Uses the map metaphor and browser metaphor; consistent design, multi-modal interaction; hierarchy of maps (overview maps in raster and detailed maps in vector); high contrast for outdoor use; POIs grouped into regions to avoid clutter on overview maps; speech commands; users views, not application modes; multi-lingual (English, German, French); graphical, text, and voice routing	
BUSINESS/REVENUE MODEL: Charged through phone bill; additional revenue for tourist agencies from access to tour diary and production of diary on CD/DVD (Umlauft et al., 2003)	
PERFORMANCE/PROBLEMS/ISSUES: Network and device constraints; user in a strange environment; must work 'out of the box'; connection loss addressed through seamless restart; network-initiated push mechanism to reduce network traffic; part of business logic resides in mobile device; locational accuracy and scale; automated labelling issues (use of tool tips and hotspots); use of 3D silhouettes and stretching the map to improve user orientation; interactive and dynamic mapping; assume low-level devices (limited hard keys, uses soft keys); data compression; data integration; symbology (simple, self-explanatory)	

- the use of 3D maps on mobile devices (Laakso, Gjesdal, & Sulebak, 2003);
- The Guide, which aimed to develop a context-aware tourist guide for Lancaster (England), which overcame many of the limitations of existing information and navigation tools (Cheverst et al., 2000);
 - The Electronic Guidebook, which was developed to show how wireless technology could be used to enhance all three phases of a visitor's museum experience (Semper & Spasojevic, 2002).

The development of such tourism applications can be expected to have a significant impact on

the tourism industry, as it will create a greater awareness of the tourism industry for business and leisure activities, as well as a greater appreciation of the history, culture, and environment of places. It will improve the knowledge-acquisition process of tourists, and this increased knowledge can then help conservation and preservation efforts. This in turn can be expected to lead to an increase in tourism. Hence, the revenue and economic development generated from such applications could be considerable (O'Looney, 2004, p. 14). In this way such services could be of great benefit to the tourism industry, and revenue generated from mobile tourism could be fed back into the industry.

Table 4. Webpark mobile tourist guide (Krug, Mountain, & Phan, 2003; Edwardes, Burghardt, & Weibel, 2003)

APPLICATION AREA: Rural and Recreation Tourist Guide	
B2B/B2C: B2B and B2C	IMPLEMENTED: Prototypes trialled
SUMMARY: Webpark aims to create a platform for LBS for tourists in rural, recreation, and protected areas across Europe (e.g., National Parks). Utilising existing information, Webpark will create new value-added services (and new value chains), providing location-specific, timely, and personalised information (maps, text, multimedia information from several databases) and services to cyclists, hikers, and so forth to aid their choices. Such services are enhanced through data mining (e.g., to predict travel times or accessibility). Webpark can also be used by administrators to not only collect and manage information, but also to aid in the protection of natural resources, educate, and influence attitudes towards conservation (www.webparkservices.info/pages/project.html)	
PUSH/PULL SERVICES PROVIDED: Push—safety, weather, and ecological alerts; spatial and temporal queries about flora and fauna (e.g., where are eagles found, which species are found close by at this time of year); species identification; reconnaissance; route guidance, tracking, and route profiles; information on POIs, hotels, restaurants (e.g., visible POIs filtered according to user preferences); record personal information on places visited (notes, photos, etc.) Pull—emergency services; visitor tracking (providing information to administrators on user preferences, where visitors go and when, travel times according to different modes of transport, etc.); the user is requested to opt in for visitor tracking	
GEOGRAPHIC COVERAGE/ENVIRONMENT: Developed for outdoor rural and recreation environments of Europe; trialled in the Swiss National Park, Dartmoor National Park, and Wadden Sea National Park	
POSITIONING TECHNIQUES: GPS (not restricted by network coverage); also utilises mobile communications networks; able to utilise any positioning technology	
MOBILE DEVICES SUPPORTED: Smart phones and PDAs; initial trials used a Compaq PDA (iPAQ and Navman GPS) and a Nokia GPRS phone	
COMMUNICATIONS INFRASTRUCTURE: GPRS capabilities tested in Swiss National Park trials	
DESIGN INTERFACE: Initial trials used ArcPad, which was replaced by their own platform due to problems, such as too many buttons and menus, and instability of the application (Edwardes et al., 2003, p. 1013); Webpark utilises three groups of controls—a query interface (lists and maps), explore answers interface, and display answers interface (lists, maps, multimedia information); for map interfaces, commonly used controls such as pan, zoom, and so forth are available	
BUSINESS/REVENUE MODEL: Free and pay-per-use services; micro-payments can be used from an Internet wallet account; premium services require user subscription (for additional personalisation and push data, such as weather alerts)	
PERFORMANCE/PROBLEMS/ISSUES: User needs, information availability, delivery mechanisms, interoperability; knowledge discovery, the use of intelligent agents; dynamic visualisation on small displays (e.g., scale and generalisation issues); limited network coverage in rural areas; the need for spatial and temporal metadata for value-added personalised services; data integration; privacy; security; pricing	

Such applications will also fundamentally alter the experience of the tourist, allowing the user to spend more time exploring sites and attractions, and less time finding information, making bookings, and so forth. This is an important consideration when a significant component of the tourism market is dominated by “time poor” individuals who take frequent but shorter holidays. The implementation of mobile tourist guides will also result in the creation of new and longer value chains, including network providers, a number of content providers, application developers, and so on. They will also

lead to increased specialisation (e.g., different content providers specialising in different types of information) (Peng & Tsou, 2003, pp. xxx-xxxi). Another possible benefit is improved safety for tourists.

ADDRESSING APPLICATION DEVELOPMENT CHALLENGES

While prototypes and applications have been developed and they clearly show the potential and benefits of mobile technology for the mo-

mobile tourist, there are a number of technical and business issues which must be addressed before such applications will be widely adopted. The technology problems must be solved first (i.e., the applications have to “work” before cost and other issues can be tackled). Yet studies have shown that customers are flexible, provided there are significant benefits and the costs of the new technology are lower (e.g., SMS) (Eriksson, 2002, p. 15). It should also be remembered that we are dealing with B2C applications here and that consumers require them to be easy to use, low cost, and of better value than existing products/solutions.

Technical Challenges

As mobile tourism applications depend on a relatively complex set of components, the first challenge is to ensure that the various components that make up such applications can communicate with each other. This is a significant problem given that different mobile devices, networks, GIS software, data servers, and so forth utilise different interfaces and protocols (Chen, 2004). The second challenge is to ensure that mobile tourism applications can be integrated with existing (non-mobile) systems (e.g., billing systems) (McKee, 2004; Spinney, 2003)—hence, the need for standard APIs and protocols (e.g., XML, GML, Java, OSA/Parlay API, CORBA). Such standards ensure that each application can access distributed and heterogeneous data sources, distributed tools, and services, regardless of the underlying platform, hardware, vendor, network, positioning technology, data formats, and so forth (Zadorozhny & Chrysanthis, 2004, p. 147). Standards also allow for mobile tourism applications to be easily implemented with minimal cost and effort (McKee, 2004). A number of standards have been developed for LBSs, the most significant of which are the Open GIS

Consortium’s (OGC) Open Location Service (OpenLS) GeoMobility Server (GMS) and the mobile location protocol (MLP) of the Location Interoperability Forum (LIF).

The methods of user interaction also differ from those for Internet- or PC-based applications. Small-screen mobile devices allow fewer means of user input. PDAs and smartphones utilise a touch screen and pointing device rather than a mouse; some PDAs have a qwerty keyboard, but most rely on onscreen keyboards or handwriting recognition software; they also have scrolling keys, and some support voice entry. Smartphones also utilise onscreen keyboards and scrolling keys. WAP phones, on the other hand, tend to have even more restricted means of user input, being limited to phone keypad entry and scrolling keys (Peng & Tsou, 2003, p. 456) (see Table 5). Hence, PDAs and smartphones are the preferred devices.

Such devices are also limited by their small screen size, so the user interface must fit within a very small area. Menus, buttons, tabs, lists, maps, and so on must all be arranged so as to optimise the limited space available. Each application must be customised in order to provide a clear, uncluttered, easy-to-read, and easy-to-use user interface. Most applications adopt the Web metaphor for their user interfaces. The standardisation of interfaces (e.g., icons and symbols) (Eriksson, 2002) for the wireless Web would be a significant advance in the development of mobile tourism applications.

Maps and graphics should also be designed, especially for such small-screen devices—they should be simple, uncluttered, have high contrasting colours (so they can be read outdoors in bright light), and be easy to read at arm’s length (Hjelm, 2002, p. 247). Maps should also take into account cultural conventions and utilise widely understood symbology, so that they can be understood regardless of language (Nissen, Hvas, Münster-Swendsen, & Brodersen, 2003).

Table 5. Characteristics of mobile devices (largely derived from Mountain & Raper, 2002, pp. 3-7)

Device	Description	Client-Server Model	Communication	User Interaction
WAP Phone	Largest number of users, small screen, poor resolution, sometimes monochrome but now many colour displays; narrow colour palette; mainly supports text (SMS); little processing power, closed platform; pre-installed applications, little customisation; can only operate as a connected mobile tour guide	Server side processing	Poor caching ability means can only be used online; reliant on network coverage	Phone keypad; scrolling keys
Smartphone	Larger screen size, better resolution, greater storage and memory, better colour palette and processor; can download applications, utilises common OS, e.g., Symbian; supports text, maps, e-mails, word processing, etc.; allows customisation; can operate as unconnected or connected mobile tour guides; greater functionality; fewer users	Server-side and client-side processing	Better caching capability means can be used online, as well as off-line when network coverage is unavailable	On-screen keyboard, touch screen and pointer; scrolling keys; phone keypad
PDA	Can operate as unconnected or connected mobile guides; they have the advantage of an even larger screen, higher resolution, larger colour palette, higher processing power, larger storage and memory; customisation, can download applications; Palm OS, Windows CE, etc.; supports greater client-side processing; greater functionality; higher-end applications; fewer users; supports a wider range of positioning techniques	Support the widest range of client-server models	Even better caching capability means can be used online, as well as off-line when network coverage is unavailable	Qwerty keyboard or onscreen keyboard and pointer; hand writing recognition; scrolling keys; voice input

Labelling is also a problem on such small-screen devices, in that automated labelling can result in cluttered and overlapping labels. In order to address this issue, labels should be scale dependent, and tool tips and hotspots should be used (Gartner & Uhlirz, 2001; Uhlirz, 2001).

Other limitations related to the use of mobile devices include poor screen resolution, limited

colour palette, short battery life, limited storage and memory, and slow processing power. Many of these limitations, however, are already less of an issue given the pace at which hardware is improving. Nevertheless, small screen size will continue to be an issue, regardless of other technological improvements.

Applications must be smart enough to be able to deliver content to the user, taking into

account the user device and its capabilities, as well as the characteristics of the wireless network. For example, owing to the limitations of WAP phones, it is preferable to deliver driving directions to these devices as text rather than as maps and multimedia information; the same query on a PDA may result in the delivery of maps and audio to the driver.

Web sites developed for the wired Internet are generally unsuitable for small-screen mobile devices and must be redesigned for these mobile devices—shorter pages, a deeper hierarchy, and so forth (Kacin, n.d.). Furthermore, the information delivered to such devices must be in a form the mobile devices can read (e.g., WML, WAP, HDML, C-HTML, XHTML) (Peng & Tsou, 2003, p. 468). The format of spatial data delivered to the Internet is also inappropriate to the wireless Web (e.g., GML, SVG) (De Vita, Piras, & Sanna, 2003; Reichenbacher 2003). Hence, data formats suitable for wireless devices have been developed. Most applications using vector data use GML to store the data on the server and convert to it to SVG Tiny (for mobile phones) or SVG Basic (for PDAs) for the display of vector data (e.g., Reichenbacher, 2001a, 2001b, 2003; Nissen et al., 2003).

Many applications also utilise predefined, static, scanned raster maps, as they result in lower storage, transmission, and rendering costs (Pospischil et al., 2002). The problem is, however, that they allow for little manipulation, interactive, or adaptive mapping, and rescaling results in information loss (Reichenbacher, 2001a, 2001b, 2003; Hjelm, 2002, p. 257). Raster data formats also result in high storage and transmission costs. Vector data formats, on the other hand, result in lower storage and transmission costs, but greater rendering costs (Hjelm, 2002, p. 257). However, they do support dynamic, interactive, and context-adaptive mapping, connection to a database, and a wider

range of applications (Abowd et al., 1997; Long, Kooper, Abowd, & Atkeson, 1996; Reichenbacher 2001a, 2001b).

The use of wireless networks also poses a challenge for mobile tourism. The limited bandwidth of wireless networks is a particular issue for the delivery of spatial data and multimedia content. As applications for mobile tourists require real-time delivery of information, users will not take up these services if they have to wait (i.e., the issue of latency). Hence, such applications require reasonably fast transmission and rendering of spatial data. Other issues particularly relevant to telecommunications networks are connection stability and availability (Peng & Tsou, 2003, p. 467). Coverage is also a major issue for wireless networks, as without access to the network, no information can be exchanged with the server. In this case, the applications must rely on locally cached data. Given the limited memory and storage of many mobile devices, this is a problem. Some wireless networks, such as WLAN and Bluetooth, also only cover small geographical areas.

Given that mobile phones are the most common form of mobile devices for leisure tourists, another important issue is that of international roaming and whether the mobile device can connect to the “local” mobile network. For example, a GPRS phone, operating in a GSM network environment, can only operate in GSM mode. If a GSM phone is being used in North America where the network is AMPS, then the phone cannot connect to the local network. With the introduction of GSM networks in North America, this is becoming less of a problem, although tourists from GSM countries will need to own a phone with the ability to access the GSM1900 band (these units are popularly referred to as “tri-band” phones; “quad-band” phones, which can also access GSM850, are available now).

Another major issue for mobile tourism is that of locating the user at the required accuracy. Information delivered to the tourist is location dependent, so this is an important aspect of such applications. Not only are there a wide variety of positioning technologies available (mobile network based and handset based), but they each have different limitations and accuracy. GPS, a handset-based technique, is generally regarded as superior, as it provides global coverage and is more accurate (around 10m) than network-based techniques, but GPS has operational limitations in urban environments and indoors.

Mobile network-based positioning does not require any direct expenditure by the user, as existing mobile phone networks already have the ability to identify the user's location by Cell ID. More precise positioning techniques, however, require upgrades to the network infrastructure (e.g., Time of Arrival—TOA). Some, such as Enhanced Time Difference of Arrival (EOTD), also require more intelligent handsets. As such techniques require no additional outlay by the user (other than the cost of their mobile device) and can be utilised by even the simplest devices, they are the easiest positioning techniques to use for consumer applications. However, network-based techniques, such as Cell ID, TOA, and so on, have a very coarse accuracy (50-500m in urban areas), which can be a major problem for tourists trying to navigate around a city.

GPS positioning is the better choice for mobile tourism applications, but requires the mobile device to have a built-in GPS receiver or to be able to connect to a GPS receiver—hence, more cost for the tourist. While PDAs with built-in GPS receivers are on the market, they are costly and few mobile phones today come with in-built GPS receivers. Nor are mobile tourists likely to want to carry around two devices (a GPS receiver and a mobile device).

Business Challenges

Revenue/business models for mobile tourism will vary according to the application/services provided (Giaglis, Kourouthanassis, & Tsamakos, 2002, p. 76). They can, however, be quite complex, owing to the longer value chain involved in such services. Hence, while it may be that the network providers may charge the users directly, other stakeholders, such as content providers, will want a share of the revenue (Giaglis et al., 2002, p. 74). Thus the stakeholders will have to agree on costs, as well as on issues such as roaming and privacy protocols (Eriksson, 2003, p. 17). A centralised billing system in which network providers collect fees and then forward them on to other stakeholders is easiest for both tourists and providers alike (Watson et al., 2004, p. 324).

There are a number of business/revenue models adopted by existing applications with the basic distinction being whether the service is free of charge or the user is charged (Giaglis et al., 2002). In some cases only certain services are free, while charges are incurred for higher-level services (e.g., Webpark). Charging may be by subscription or on a pay-per-use basis. If the pay-per-use model is used, there must be a system in use for micropayments. Giaglis et al. (2002, pp. 76-78) offer a framework for matching services/applications, technology, and business models, in order to assist in the development of such applications.

Fundamentally, there are three sets of charges: (1) cost of handsets (purchase or rental); (2) mobile telephone infrastructure costs (rental of the sites the base stations are situated on, the cost of the base station equipment, ongoing maintenance costs, access costs to the sites, roaming charges); and (3) charging for value-added services. O'Looney regards syndication as an effective business model for such services, owing to the number of stakeholders

involved. Syndication, however, requires standardisation of information and transaction procedures (O’Looney, 2004, p. 4).

The costs of mobile devices are also an issue, as are the costs of the services themselves. Until the cost of suitable mobile devices is more affordable to leisure travellers, mobile tourism applications will not be widely adopted (although the current popularity of “multi-function” phones—which include some PDA functionality, camera, and MP3 players—suggests that consumer resistance may be overstated). At the same time, tourists have to be willing to pay for such services. In order to gain acceptance, mobile tourism applications must offer real-time, value-added, accurate, location-specific services that provide significant advantages over traditional modes of tourism. The availability of suitable and accurate content (data quality) is another factor inhibiting the widespread adoption of such services—some argue that suitable content will only become available when there is a critical mass of mobile tourism applications in the marketplace. The debate can be summarised simply as one of whether “applications availability drives content” or “content availability drives application development.” In order to ensure the widespread development of mobile tourism, Eriksson (2002, p. 17) states that the tourism industry needs to: store all information digitally and make it accessible via digital devices; all services must be able to be paid for digitally; information must be made available in a variety of languages; and all facilities must be able to be plotted on maps.

A major factor limiting consumer acceptance is the issue of privacy/security. These applications store vast amounts of personal information on users (e.g., user profiles, user location, etc.), and service providers need to protect this sensitive information by implementing clear security and privacy protocols; secure

encryption during transmission is a minimum requirement. Consumers need to know that service providers will not use this information in any other way or pass on their personal details to any third party. Many applications address this issue by requiring the user to opt-in or opt-out of services (e.g., user tracking). This also solves the problem of some services becoming a nuisance (e.g., marketing), as they can be turned off. Another way to ensure trust is to make sure control of the location information stays in the hands of the consumer (Giaglis et al., 2002, p. 78). Until consumers trust the service providers, such applications will not be widely adopted.

FUTURE DIRECTIONS

The adoption and implementation of standards for mobile tourism will assure “plug and play” capability of LBS applications on different devices or different telecommunications networks and across different platforms. With technological improvements in hardware and wireless networks (especially with the adoption of 3G or better networks), many of the technical challenges mentioned above will no longer be an issue. Along with these improvements, other features will become more prominent, such as 3D visualisation, virtual reality, and animations. Applications are already being developed with 3D visualisation, virtual tours, and animations as their focus (e.g., TellMaris and DeepMap).

The rapid rate at which mobile technology is improving and the decreasing costs of mobile devices will mean that suitable and cost-effective devices will soon be available to a wider range of consumers. Currently the cost of fully featured smartphones and PDAs is too high to lead to widespread use among leisure travellers, although there is widespread use of these devices among business tourists.

Technological improvements in positioning techniques will also result in improved positional accuracy (e.g., as a result of improvements in networks—3G networks, Galileo, etc.). The decreasing costs and shrinking size of GPS receivers and the further development of mobile phones with inbuilt GPS receivers (early versions of these units are available) will also mean that GPS positioning will be more widely available. GPS provides more accurate positioning (than those based on telecommunications networks), as well as global coverage. Hence, it is more appropriate for providing accurate and value-added services to the mobile tourist. Even when mobile networks are unavailable, real-time positioning is still possible. While more precise non-GPS-based positioning is available using WLAN, Bluetooth, and so forth, they have very restricted local coverage.

A significant restriction to the widespread adoption of mobile technologies for tourism is the unavailability of suitable content. As more applications are developed and more content providers become aware of the potential of mobile tourism, more content will become available. This will also result in increased specialisation of the content providers.

CONCLUSION

Mobile tourism has the potential to enhance the experience of the mobile traveller and to deliver new kinds of information to the user—animations, video, audio. The most significant advantage to the tourist is to not only support the pre- and post-touring phases of activity, but also to support the mobile tourist on the move. By providing value-added, location-specific, personalised services to mobile tourists, mobile technologies will not only offer superior services to the travellers (services not provided by

existing Web-based applications), but also fulfil a requirement of the mobile user.

Services that improve the efficiency and experience of the tourist are needed more than ever before, in an environment where people have less leisure time and shorter holidays. Such services will, however, not be adopted unless they provide significant advantages at reduced costs. A number of technical and business challenges must also be addressed before applications for the mobile tourist will be widely adopted. With the rapid speed at which technology is improving, a major market is likely to emerge in the next few years.

ENDNOTES

- ¹ Such applications fall under the banner of location-based services (LBSs), which refer to a suite of applications which utilise the location of a mobile device in order to deliver real-time, location-specific, value-added, personalised services.
- ² The tourism industry could certainly benefit from a standard data model or interface from which to access structured data, so enhancing user satisfaction with the planning stage of the tourism activity (Watson et al., 2004, p. 316).

REFERENCES

- Abowd, G., Atkeson, C., Hong, J., Long, S., Kooper, R., & Pinkerton, M. (1997). Cyberguide: A mobile context-aware tour guide. *Wireless Networks*, 3, 421-433.
- Annegg, H., Kunczier, H., Michlmayr, E., Pospischil, G., & Umlauf, M. (2002). LOL@: Designing a location based UMTS application. *Elektrotechnik und Informationstechnik*, 119(2), 48-51.

- Brown, B., & Chalmers, M. (2003, September 14-18). Tourism and mobile technology. In K. Kuutti, E. Karsten, G. Fitzpatrick, P. Dourish, & K. Schmidt (Eds.), *Proceedings of the 8th European Conference on Computer Supported Cooperative Work*, Helsinki, Finland (pp. 335-355). Dordrecht: Kluwer Academic Press.
- Buhalis, D. (2003). *E-tourism: Information technology for tourism management*. London: Prentice-Hall.
- Chen, A. (2004). *Open standards will evolve location-based services*. Retrieved November 30, 2004, from http://www.eweek.com/print_article2/0,2533,a=131075,00.asp
- Cheverst, K., Davies, N., Mitchell, K., Friday, A., & Efstratiou C. (2000). Developing a context-aware electronic tourist guide: Some issues and experiences. In *Proceedings of CHI2000, the Conference on Human Factors in Computing Systems* (pp. 17-24). New York: ACM Press.
- CSTB (Computer Science and Telecommunications Board). (2003). *IT roadmap to a geospatial future*. Washington, DC: The National Academies Press.
- De Vita, E., Piras, A., & Sanna, S. (2003). *Using compact GML to deploy interactive maps in mobile devices*. Retrieved from <http://www20003.org/cdrom/papers/poster/p051/p51-devita.html>
- Edwardes, A., Burghardt, D., & Weibel, R. (2003, August 10-16). Webpark—location based services for species search in recreation area. In *Proceedings of the 21st International Cartographic Conference, "Cartographic Renaissance"*, Durban, South Africa.
- Epstein, M., Garcia, C., & dal Fiore, P. (2003). *History unwired: Venice frontiers—mobile technology for intelligent tourism and citizenship*. Cambridge, MA: MIT. Retrieved from <http://web.mit.edu/frontiers>
- Eriksson, O. (2002). Location based destination information for the mobile tourist. In *Information and Communication Technologies in Tourism—2002* (pp. 255-264). New York: Springer.
- Gartner, G., & Uhlirz, S. (2001). Cartographic concepts for realising a location based UMTS service: Vienna city Guide "LOL@." In *Proceedings of the 20th International Cartographic Conference* (vol. III, S.3229-3239), Beijing. Retrieved from http://lola.ftw.at/homepage/content/a40material/Vienna_City_Guide_LoLa.pdf
- Giaglis, G., Kourouthanassis, P., & Tsamakos, A. (2002). Towards a classification for mobile location services. In B. Mennecke, & T. Strader (Eds.), *Mobile commerce: Technology, theory, and applications* (pp. 64-81). Hershey, PA: Idea Group Publishing.
- Hassin, B. (2003). *Mobile GIS: How to get there from here*. Retrieved March 29, 2005, from <http://gis.esri.com/library/userconf/proc03/p0988.pdf>
- Hjelm, J. (2002). *Creating location services for the wireless Web: Professional developer's guide*. New York: John Wiley & Sons.
- Kacin, M. (n.d.). *Optimizing Web content for handheld devices*. Retrieved from <http://www.wirelessdevnet.com/channels/pda/features/handheldcontent.html>
- Karagiozidis, M., Zacharopoulos, I., Xenakis, D., Demiris, A., & Ioannidis, N. (2003). *Location aware visually enhanced ubiquitous services*. Retrieved from http://loveus.intranet.gr/docs/LoVEUS_TechPaper.pdf

- Karimi, H., & Hammad, A. (2004). *Telegeoinformatics: Location-based computing and services*. New York: CRC Press.
- Krug, K., Mountain, D., & Phan, D. (2003, March). Webpark. Location-based services for mobile users in protected areas. *GeoInformatics*, 6(3), 26-29.
- Laakso, K., Gjesdal, O., & Sulebak, J. (2003, September 8-11). Tourist information and navigation support using 3D maps displayed on mobile devices. In B. Schmidt-Belz & K. Cheverst (Eds.), *Proceedings of the Workshop on Mobile Guides, Mobile HCI 2003 Symposium*, Udine, Italy (pp. 34-39).
- Lohnert, E., Mundle, H., Wittmann, E., & Heinrichs, G. (2004). Wireless in the Alps: An LBS prototype for mountain hikers. *GPS World*, 15(3), 30-37.
- Lohnert, E., Wittmann, E., Pielmeier, J., & Sayda, F. (2001, September 11-14). PARAMOUNT: Public Safety & Commercial Information Mobility Applications & Services in the Mountains. In *Proceedings of the 14th International Technical Meeting of the Satellite Division of the Institute of Navigation (ION GPS 2001)*, Salt Lake City, UT (pp. 319-325).
- LOL@ Local Location Assistant. (2004). Retrieved November 25, 2004, from <http://lola.ftw.at/homepage/>
- Long, S., Kooper, R., Abowd, G., & Atkeson, C. (1996, November). Rapid prototyping of mobile context-aware applications: The Cyberguide case study. In *Proceedings of the 2nd ACM International Conference on Mobile Computing and Networking (MobiCom '96)*, Rye, NY (pp. 97-107). New York: ACM Press.
- LOVEUS. (2004). Retrieved December 5, 2004, from <http://loveus.intranet.gr/>
- McKee, L. (2004). LBS interoperability through standards. In J. Schiller & A. Voisard (Eds.), *Location-based services* (pp. 149-171). Amsterdam: Elsevier.
- Mountain, D., & Raper, J. (2002, September 18). Location-based services in remote areas. In *Proceedings of the Association of Geographical Information* (Paper B5.1, pp. 1-9). Retrieved from <http://www.soi.city.ac.uk/~dmm/research/pubs/B05.3.pdf>
- Nissen, F., Hvas, A., Münster-Swendsen, J., & Brodersen, L. (2003). *KMS, National Survey and Cadastre—Denmark: Small-display cartography*. Retrieved November 30, 2004, from http://gimodig.fgi.fi/pub_deliverables/D3_1_1.pdf
- O'Looney, J. (2004). GIS and enlightened location-based tourism: An innovation whose time has come. In *Proceedings of the ESRI User Conference*, San Diego, CA.
- Open LS. (2004). *Open GIS location services*. Retrieved December 15, 2004, from <http://www.opengeospatial.org/specs/?page=specs>
- PARAMOUNT. (n.d.). Retrieved from <http://www.paramount-tours.com>
- Peng, Z., & Tsou, M. (2003). *Internet GIS: Distributed geographic information services for the Internet and wireless networks*. New York: John Wiley & Sons.
- Pospischil, G., Umlauf, M., & Michlmayr, E. (2002). Designing LOL@, a mobile tourist guide for UMTS. *Lecture Notes in Computer Science*, 2411, 140-154.
- Readman, D., & Mojarrabi, B. (2004). *Location based and communication systems*. Unpublished manuscript, University of Queensland, Australia.

- Reichenbacher, T. (2001a). Adaptive concepts for a mobile cartography. *Supplement Journal of Geographical Sciences*, 11, 43-53.
- Reichenbacher, T. (2001b). The world in your pocket—Towards a mobile cartography. In *Proceedings of the 20th International Cartographic Conference*, Beijing, China. Retrieved from http://citeseer.ist.psu.edu/cache/papers/cs/23234/http:zSzzSzwww.lrz-muenchen.de:zSz~t583101zSzWWWzSzpublicationszSzreichenbacherzSzICC2001_Paper.pdf/reichenbacher01world.pdf
- Reichenbacher, T. (2003, August 10-16). Adaptive methods for mobile cartography. In *Proceedings of the 21st International Cartographic Conference (ICC)* (pp. 1311-1322), Durban, South Africa. Retrieved from http://www.carto.net/geog234/readings/reichenbacher_mobile_cartography_durban_2003.pdf
- Semper, R., & Spasojevic, M. (2002, April). The electronic guidebook: Using portable devices and a wireless Web-based network to extend the museum experience. In *Proceedings of the Museums and the Web Conference*. Retrieved from <http://www.archimuse.com/mw2002/papers/semper/semper.html>
- Spinney, J. (2003). *A brief history of LBS and how OpenLS fits into the new value chain*. Retrieved from http://www.esbic.ie/geo-business/Mobile_GIS/overview.htm
- TellMaris. (2004). Retrieved December 10, 2004, from <http://www.tellmaris.com/>
- The Electronic Guidebook*. (2005). Retrieved April 4, 2005, from <http://www.exploratorium.edu/guidebook/>
- Uhlirz, S. (2001). Cartographic concepts for UMTS-location based services. In *Proceedings of the 3rd Workshop on Mobile Mapping Technology*, Cairo, Egypt. Retrieved from http://lola.ftw.at/homepage/content/a40-material/Cartographic_Concepts_for_UMTS_Location_based_Services.pdf
- Umlauf, M., Pospischil, G., Niklfeld, G., & Michlmayr, E. (2003). LOL@, a mobile tourist guide for UMTS. In H. Werthner, & E. Veit (Eds.), *Journal of Information Technology & Tourism*, 5(3), 151-64.
- Watson, R., Akselsen, S., Monod, E., & Pitt, L. (2004). The Open Tourism Consortium: Laying the foundations for the future of tourism. *European Management Journal*, 22(3), 315-326.
- Webpark Mobile Tourist Guide. (2004). Retrieved December 10, 2004, from <http://www.webparkservices.info/>
- Wherify. (2005). Retrieved March 24, 2005, from <http://www.wherifywireless.com/univLoc.asp>
- WTTC. (2003). *Travel and tourism: A world of opportunity*. Retrieved from [http://www.wttc.org/measure/PDF/Executive% 20 Summary.pdf](http://www.wttc.org/measure/PDF/Executive%20Summary.pdf)
- Zadorozhny, V., & Chrysanthi, P. (2004). Location-based computing. In H. Karimi & A. Hammad (Eds.), *Telegeoinformatics: Location-based computing and services* (pp. 145-170). Boca Raton, FL: CRC Press.

Chapter LIX

Mobile Computing—An Enabler in International Financial Services

N. Raghavendra Rao

SSN School of Management & Computer Applications, India

ABSTRACT

Information and telecommunication technologies are the major stimulus for changes in trade and commerce. Recent convergence of the above technologies has become possible due to the rapid advancements made in the respective technology. This convergence is termed as information and communication technology (ICT) and considered as a new discipline. The new discipline has made cross border commerce in the present globalization scenario a reality. This chapter talks about a model for financial services sector in international market under the new discipline. The model explains the creation of knowledge based financial services system incorporating the sophisticated concepts of information technology. Further, it provides an access to the system with devices which can be used under wireless communication environment, across the globe.

INTRODUCTION

The effects of the convergence of telecommunication and information technology are being felt in the present global corporate world. This new discipline has made economics across the globe closely interconnected and integrated. Business processes are constantly changing at an exponential rate. The new discipline is also advancing by delivering exponential increase in

computing power and communication capability. The result of this advancement has created a new generation of computers working on wireless technology, cell phones having the features of portable computers, and notebooks offering similar performance of desktop computers by using the same software. Portable computers and cell phones are no longer just for globetrotting executives. Innovations and radical changes are taking place in these products.

The approach of the makers of these products is to provide fast and unwired connections in their products, enabling their clients to make use of the rich resources of their organizations located across the globe.

The policy of globalization followed by many countries is changing the world's financial markets. In this context, Buckley (2003) observes that the world economy is internationalizing and, further, firms may engage in the international business by undertaking portfolio investment (p. 35).

This has led to deregulation. This is also providing opportunities to many financial institutions across the globe who are rendering investment advisory services. Accordingly every country is rapidly adapting itself to the new global changing vistas in the financial market. It is high time the investment advisory service providers take advantage of the benefits from the new discipline. A model is suggested to help investment advisors who are involved in the international financial market analyzing data and information for investment. Further it provides information to their team members who are located at various locations across the globe for providing services.

Business Process

The international financial market mainly comprises the corporate securities, Forex, metals, and commodity segments. Investment decision and advice in these segments need vast information. Information is required for corporate companies regarding the industry, natural resources such as metals, commodities, and the country level of each segment. The types of databases which can hold a high volume of data and information are required for this model. Sophisticated software tools are also needed for analyzing the data and information from these databases.

Investment financial analysts often explore an incredible amount of data about instruments, markets, and the corporate sector. They analyze the different market segments, price movements, economic forecasts, and news events. They react on the basis of market information, price trends, historical data, and their own experience. In this process, they can make many observations from the data and information available. They can try to determine the patterns from their observations.

Case Study for International Financial Services

A London-based investment consultancy organization, which has been operating in securities trading at the London Exchange market, has decided to go global. The organization decided to add other activities such as securities related to companies in different countries, Forex, metals, and commodities as their core services under its umbrella. It also changed its name to the Global Finance Services Advisory Group (GFSAG). GFSAG hired domain experts located in different countries under its business process outsourcing strategy. The group decided to follow the concept of virtual office for its operations in different countries. Domain experts and their team members can operate from anyplace of convenience. Their approach for virtual offices is to save the cost of infrastructure and to avail the benefits under the new discipline. The respective domain expert groups are expected to monitor, guide, and assist their counterparts and team members at different locations across the globe. The corporate office in London provides services for all activities to existing clients and prospective clients through the executives located at various locations across the world. In case of additional information and clarification, the executives are permitted to be in touch with the

Table 1. Activities of GFSAG

Country	Location	Activities	Controlled By
USA	New York	Corporate Securities	Domain Experts
Australia	Sydney	Foreign Exchange	Domain Experts
Middle East	Bahrain	Metals	Domain Experts
Japan	Tokyo	Commodities	Domain Experts
UK	London	All the Above Activities	Corporate Group

respective domain experts while they are at their clients' offices. The activities of GFSAG are summarized in Table 1. The places of operations are assumed for the purpose of case study.

These domain experts can analyze the information from the knowledge-based system for forecasting and identifying the risks associated with the operations in investments in the international financial market. On the basis of their analysis, inferences can be drawn and solutions can be suggested by them. These solutions are stored in an application database in a mobile computing server. The executives at the respective locations of their offices across the globe will be guiding their clients by having access to this server.

MODEL FOR GLOBAL FINANCIAL SERVICES ADVISORY GROUP

The business process explained in the case study will be the base for creation of knowledge-based international financial services systems under a wireless communication environment. This model will be referred as the GFSAG model. The GFSAG model has the following four stages:

- **Step 1:** Creation of Knowledge-Based System in GFSAG Model
- **Step 2:** Simulation and Forecasting for the Probable Risks in Financial Market

- **Step 3:** Mobile Computing Function in GFSAG Model
- **Step 4:** Requirements for GFSAG Model

Knowledge-Based System

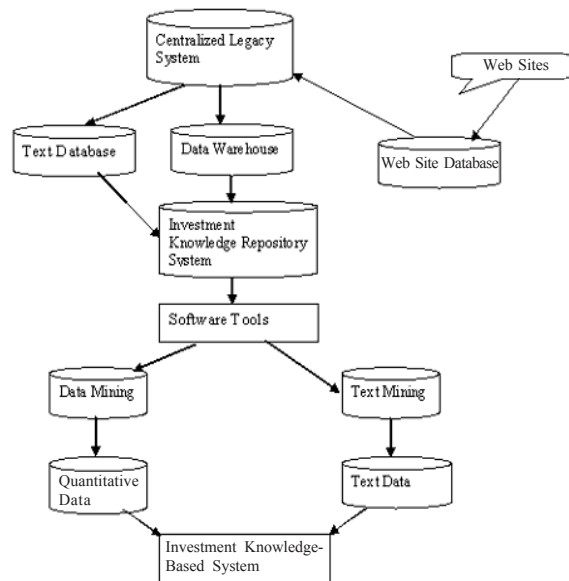
Step 1: Creation of Knowledge-Based System in GFSAG Model

A core team of GFSAG will consist of the domain experts, hardcore software professionals, and telecommunication experts. The importance of the services of hardcore software professionals and telecommunications experts cannot be underestimated because they are the backbone of the knowledge-based and core team. The macro-level design of an investment knowledge-based system is described in Figure 1.

The inputs received from the respective domain teams and other related information are stored in the centralized legacy system transferring to text database, and data warehouse depends on the type of data and information required for the creation of an investment knowledge repository system. The importance of a data warehouse in the financial sector is best described by Humphries, Hawkins, and Dy (1999):

A data warehouse contains data extracted from the many operational systems of the enterprise, possibly supplemented by external data. For example, a typical banking data warehouse will require the integration of

Figure 1. Knowledge base in GFSAG model



data drawn from the deposit systems, loan systems and the general ledger, just to name three. (p. 34)

The significance of a data warehouse is also highlighted by Adriaans and Zantinge (1999) when they say:

In order to perform any trend analysis you must have access to all the information needed to support you and this information stored in large data bases. The easier way to gain access to this data and facilitate effective decision making is to set up a data warehouse. (p. 25)

Text Database

This database will contain business practices, procedures, policies, culture, legal, taxation, accounting standards, political environment, various organizations profiles, information pertaining to natural resources of metals and commodities, and views and opinions of domain experts at each country and global level.

Data Warehouse

This will contain the quantitative data related to corporate securities, foreign exchange, metals, and commodities at each country and global level.

Web Site Database

This will contain the downloaded relevant information from the various sites in respect to the financial services sector at each country and global level.

Knowledge Repository

The data and information in the text database and data warehouse are to be grouped and stored as per the segments of the business activities of GFSAG.

Software Tools

The analysis of quantitative data stored in the investment knowledge repository system is carried through data mining. This tool helps one to know the relationship and patterns between data elements. The analysis of textual data is carried through text mining like data mining; it helps to identify relationships among the vast amount of text data. Pujari (2002) also states that text mining corresponds to the extension of the data mining approach to textual data (p. 239).

The investment knowledge-based system is created after the analysis by domain experts. This will contain how financial markets react to an event and the behavior of the market in the recent times.

Step 2: Simulation and Forecasting for the Probable Risks in the Financial Market

Virtual reality is one of the concepts among the number of other concepts provided by information technology. Simulation is the basic element in virtual reality applications. The time dividing

between simulated tasks and their real-world counterparts is very thin.

The synergy between real-world and simulated facts yields a surprising amount of effectiveness. The possible financial risks can be envisaged through simulations from the investment knowledge-based system. The simulated information can be provided through a mobile computing system. The use of features of virtual reality for simulation and forecasting for probable risks in financial markets are illustrated in Figure 2. Chorafas and Steinmann (1995) observe that each financial institution has a different way of looking at the market and business opportunity. The strategic approach must be mapped into the machine and then interactively visualized. Not surprisingly, some banks are very advanced and are leaving their competitors in the dust (pp. 174-175).

The harsh realities of risks are well known and understood by domain experts and their team members. The sophisticated tools help financial analysts form their views and opinions. It must be remembered that these tools are useful for keeping the unpleasant surprises to a minimum. The culture of the country influences the percentage of risks one takes.

Wireless Environment: Concept of Mobile Computing

Mobile computing can be defined as a computing environment over a physical mobility. Schiller (2004) rightly says that GSM (Global System for Mobile communication) is the most successful digital mobile telecommunication system in the world today (p. 96).

The main features of a GSM system are indicated in Figure 3. A GSM has three subsystems: RSS (Radio SubSystem), NSS (Network and Switching Subsystems), and OSS (Operation SubSystem).

Support for Mobility

A mobile computing network becomes more useful when it supports business applications on its network. Now it is becoming possible by adding components such as file systems, databases, and security in mobile and wireless communication. The Web has been designed for conventional computers and fixed networks. Several new system architectures offer the opportunity to change the phase in telecommunication technology. Mobile communication is being influenced by merging telecommunication with computer networks. The present trend

Figure 2. Virtual reality concepts in GFSAG model

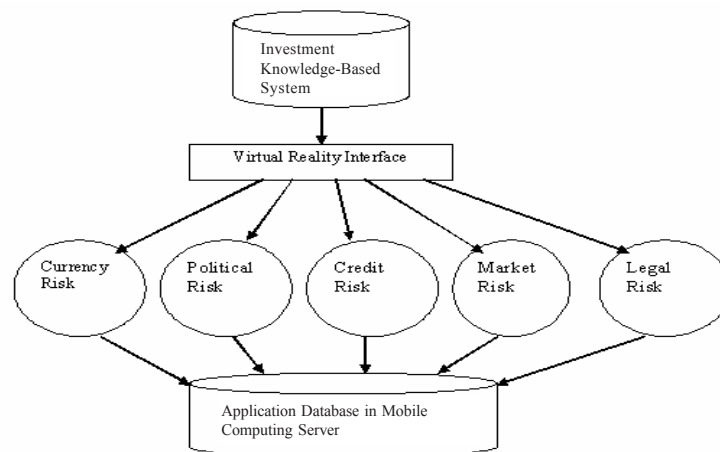
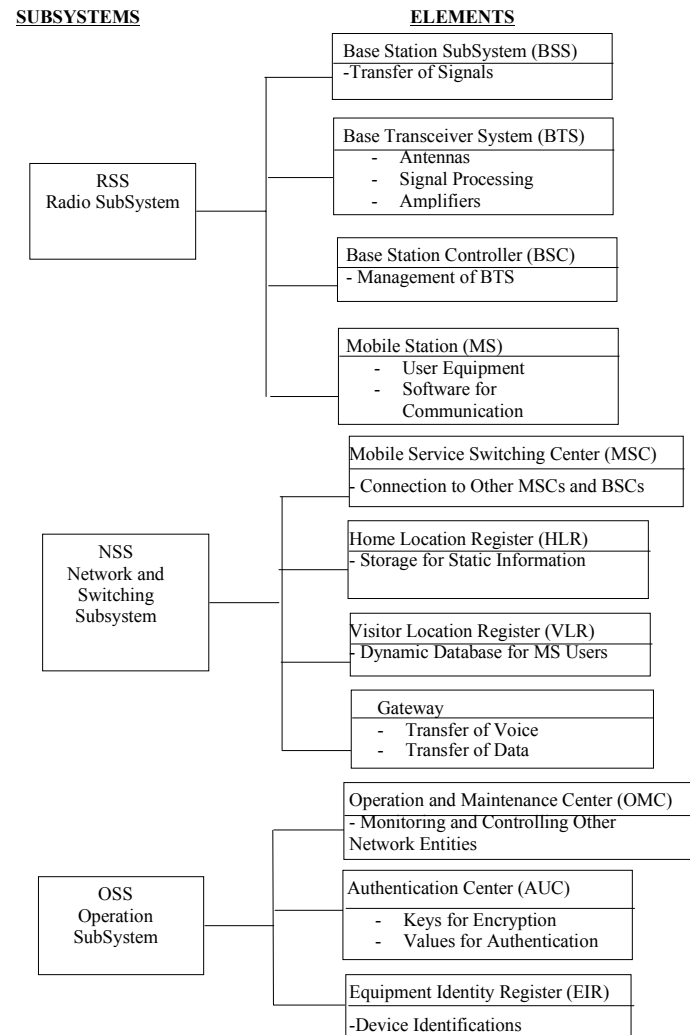


Figure 3. Overview of GSM system (adapted from Schiller, 2004)



in the mobile phone market is cell phones being designed that take care of some features of computers besides voice calls. It would be apt to refer to the present mobile phone as a “mobile device” because these have additional features besides the conventional cell phones. It is interesting to note the observation of Giussani (2001) on mobile phones (pp. 227-247). He classifies it into four categories of devices: (1) dedicated devices, (2) integrated devices, (3) modular devices, and (4) federated

devices. The essential features of the four categories of devices are mentioned in the Table 2.

Present Scenario

Recently, handset makers have been queuing up to launch a new generation of smart phones which provide many features that have been special to pocket PC and PDAs, showing it is a buyer’s market. Dood (2003) rightly points out that competition has benefited customers by

Table 2. Categories and the features of devices

Categories of Devices	Features of Devices
Dedicated	Designed for a particular functionalities
Integrated Devices	Integration of functions of different devices
Modular Devices	Bringing devices into one shell
Federated	Connecting different parts of devices

triggering price decreases and wider availability of service (p. 387).

Makers of pocket PCs are enhancing many new features in their new models. Here it would be apt to quote Dornan (2001), who states:

The hype surrounding mobile data is ultimately founded on one thing: The Internet, Vendors and Operators alike use slogan such as 'Internet anywhere' and 'Internet in your pocket', promising to cut the Internet free from its PC-based roots. (p. 190)

In the case of smart phones, the additional features indicated are sending and receiving mail, Excel spreadsheets, PowerPoint presentations, and PDF files. PDAs allow the users to go online even while enjoying the usual office tools like Word, Excel, and Internet browser. A camera phone is used to capture visual information such as a phone number on a billboard instead of looking for a paper or pen to jot down the number. In the work environment some workers take pictures of finished projects to secure a visual record of completed work in case management requests such a record. On the same lines of handsets, the manufactures of notebooks and laptop computers are launching their products with wireless technology. These

products have convenient mobility available with modem, integrated LAN, or wireless connections with desktop power. Now it has become a necessity to establish synergy between mobile computing and knowledge-based business systems through these sophisticated devices. Taulkder (2002) confirms this view by saying:

Mobile computing not only offers instant information to a mobile worker; to a mobile worker it is indeed a productivity tool. Further the list of possible mobile applications can never be complete. (p. 18)

Security

The international financial services knowledge-based system is more sensitive and critical. Encryption is a solution that ensures the data content is not altered during the transmission between originator and recipient in a wireless environment. This is elaborated by Minoli and Minoli (1999), who observe:

Cipher technique[s] lend themselves more readily to automated. Theses technique[s] are uses [sic] in contemporary security tools and there are three kinds of cryptographic functions, such as Hash functions, Secret key and Public key. (p. 215)

Figure 4. Components of asymmetric crypto system

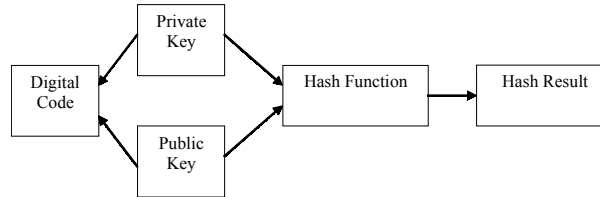
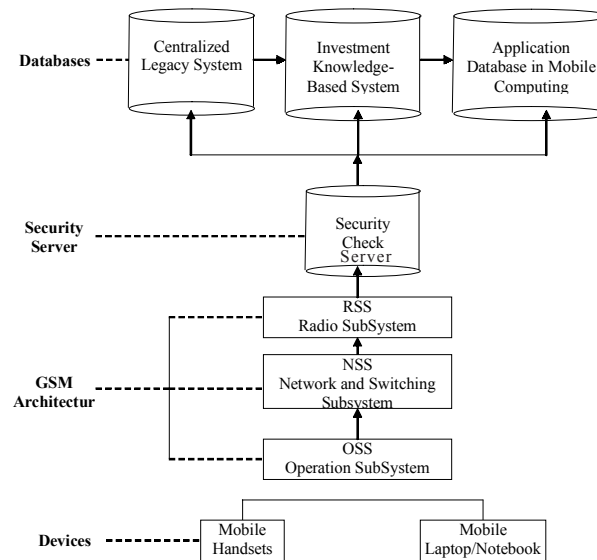


Figure 5. Mobile computing functions in GFSAG model



The concept of an asymmetric crypto system for encryption of data may be used in a wireless environment. An asymmetric crypto system is a system of secure key pair consisting of private key for creating a digital code and public key to verify the same. Hash function in this system means obtaining “hash result” by applying a predefined logic, control, or arithmetical process. Hash result means that every time the predefined procedure is applied, it should give the same result. The components in an asymmetric crypto system, taking the results obtained from the hash results, are explained in Figure 4.

With the advancement of information technology, the concept of cryptography used earlier by kings for secret communication is becoming popular in global commercial applications.

Step 3: Mobile Computing Function in GFSAG Model

The integration of the functions in mobile computing with a GFSAG model are illustrated in Figure 5.

Step 4: Requirements for GFSAG Model

The user groups for the GFSAG model at the respective country level and end users across the globe will be making use of it. The requirements of hardware, software, and mobile devices for the GFSAG model are mentioned in Table 3.

CONCLUSION

The disintegration of barriers in previously protected and insulated markets has created a new era of competition in the global economic environment. The present challenge for global players in international financial markets is how they should take advantage of the opportunities from the severe competition and survive in the market. The rewards for the opportunities are

always are accompanied by risks. Assessing risks and incorporating the same in the final decision is an integral part of the decision making. The new discipline plays an important role for acquiring and processing information for analysis and decision making. The GFSAG model offer an idea for using the services of domain experts across the globe and for minimizing the risks in financial market. The word “international” prefixed to “financial market” will become redundant once the concepts of new discipline are taken advantage of by the corporate world.

FUTURE STUDY

Many more financial markets are opening up and becoming integrated with global markets. Many virtual financial services organizations will make their presence felt in the global

Table 3. Requirements for GFSAG model (hardware, software, and mobile devices)

Particulars		Purpose
Hardware	Server	System and application programs
	Server	Exclusively for encryption and recognition of users
	Server	Storage of data and information of various activities of financial services
	Desktop/Workstation	Development of programs, and updating and retrieving data
Software	Data warehouse	Quantitative data of various segments of the financial services market
	Text Data warehouse	Text data of various segments of the financial services market
	Data mining	Analyzing data from data warehouse
	Text Mining	Analyzing text data from text data warehouse
	Virtual Reality Interface	Visualizing through simulation and applying forecasting techniques
	Other Related Software	Software for supporting the system and routine software for business purpose
Mobile Devices	Laptop/Notebook	Interaction with domain experts and corporate office Viewing the selected information from the knowledge base system Sending reports from the marketplace to corporate office and domain experts Downloading the analyzed data for understanding
	Handsets	Discreetly informing some specific information from client's place Capturing important data by using camera features in handsets, and transmitting to corporate and domain experts offices Browsing the knowledge base system for specific purpose Sending latest short news from the financial market

market. There will be an increase in reliable forecasting by using real-time data and financial modeling. Many more different activities related to financial services will be required to be added to GFSAG model.

RECOMMENDATION

The concept of virtual organization is the key to GRID computing. All the virtual organizations will share the common resources for computing power and accessing data across the globe. Grid and mobile computing concepts will be required to be integrated, once many financial markets are interconnected with each other in the global market under the concept of virtual organization. Referring to GRID computing in the financial sector, Joseph and Tein (2004) state that grid computing provides the financial analysis and services industry sector with advanced systems delivering all the competitive solutions in grid computing. These solutions exemplify the infrastructure and business agility necessary to meet and exceed the uniqueness that the financial analysis and services industry sector requires. This particular value statement is accomplished by the fact that many of these solutions in this industry are dependent upon providing increased access to massive amounts of data, real-time modeling, and faster execution by using the grid job scheduling and data access features (p. 14).

REFERENCES

- Adriaans, P., & Zantinge, D. (1999). *Data mining and data warehouse data mining* (pp. 25-36). Harlow, UK: Addison Wesley Longman.
- Buckley, A. (2003). *The internationalization process, multinational finance* (pp. 35-46). New Delhi: Prentice-Hall.
- Chorafas, D. N., & Steinmann, H. (1995). Implementing virtual reality in financial institutions. In *Virtual reality: Practical applications in business and industry* (pp. 161-179). Englewood Cliffs, NJ: Prentice-Hall.
- Dodd, A. Z. (2003). *Wireless services, the essential guide to telecommunications* (pp. 371-408). New Delhi: Pearson Education Asia.
- Dornan, A. (2001). *Inside a mobile network, the essential guide to wireless communications applications* (pp. 175-195). New Delhi: Pearson Education Asia.
- Giussani, B. (2001). *The intimate utility: Roam making sense of the wireless Internet* (pp. 227-247). London: Random House Business Books.
- Humphries, M., Hawkins, M. W., & Dy, M. C. (1999). *Data warehouse concepts, data warehousing architecture and implementation* (pp. 31-48). Englewood Cliffs, NJ: Prentice-Hall.
- Joseph, J., & Fallenstein, C. (2004). *Introduction, the grid computing anatomy* (pp. 12-14, 47-57). New Dehli: Grid Computing, Pearson Education.
- Minoli, D., & Minoli, E. (1999). *Encryption, Web commerce technology handbook* (pp. 213-225). New Delhi: Tata McGraw-Hill.
- Pujari, A. K. (2002). *Text mining, data mining techniques* (pp. 239-250). Hyderabad: Universities Press (India).
- Schiller, J. (2004). *Telecommunication systems, mobile communications* (pp. 93-130). New Delhi: Pearson Education.
- Talukder, A. K. (2002). Mobile computing—impact in our life. In C. R. Chakravarthy, L. M. Patnaik, T. Sabapathy, & M. L. Ravi (Eds.), *Harnessing and managing knowledge* (pp. 12-24). New Delhi: Tata McGraw-Hill.

ADDITIONAL READING

Cudworth, R. (2003). *The demand for continuous information: The source online business* (pp. 12-17). London: Kogan Page.

Haugen, R. A. (2002). *Securities and markets: Modern investment theory* (pp. 6-31). New Delhi: Prentice-Hall.

Lasserre, P. (2003). *Global financial management, global strategic management* (pp. 335-351). Hampshire, UK: Palgrave MacMillan.

Lumby, S. (1998). *Foreign exchange risk management, investment appraisal and financial decisions* (pp. 579-596). London: International Thomson Publishing.

Chapter LX

Mobile Computing: An Australian Case Study

Paul Hawking

Victoria University, Australia

Gina Reyes

Victoria University, Australia

Stephen Paull

Victoria University, Australia

ABSTRACT

Companies are investigating how they can extend existing business process through the implementation of mobile computing solutions. Deloitte has developed a model which can describe the evolution of mobile solutions within a corporate setting. This chapter adopts a case study approach to investigate the adoption of a mobile solution within an Australian company and classifies the implementation as per the Deloitte model.

INTRODUCTION

If we can make a \$500 handheld device to do the same thing as a \$4,000 laptop, we've saved a ton of money. (Billy Wang, Business Development Manager, Coca Cola Corporation, Mobile Planet, 2004)

Mobility, as used in the context of technology, can be described as the ability of users,

systems, or data to perform or participate in information-processing tasks without being constrained to a fixed location. Although the possible applications incorporating mobile technology have been well documented, the actual realisation of these applications has only been a recent phenomenon. There have been a number of key enablers which have facilitated this realisation (Dedo, 2004). Paavilainen (2001) suggests that the specific characteristics of

mobile devices contribute to the ease of use of these devices and the subsequent expansion of mobile markets. Convenience, instant connectivity, ability to personalise a device, and the independence of time and location are cited as the characteristics of these mobile computing units. It was been predicted that due to the rapid expansion of high-speed mobile services, by 2007, 60% of the U.S. population will receive mobile data, an increase from 2% in 2001. The Cellular Telecommunications and Internet Association (CTIA) expects that the most popular Internet access devices will be mobile and wireless technologies, surpassing PCs (Strategis Group, 2001).

Another contributing factor is the advent of enterprise systems, in particular enterprise resource planning (ERP) systems. These systems have provided the necessary infrastructure for companies to move towards “best business practice” while at the same time providing real-time access to information. This access, originally only available internally via desktop PCs, has now been extended to Web-based applications and to mobile computing devices. Many of these ERP systems have incorporated technology and scenarios to assist with the interaction with mobile devices.

IMPACT ON ORGANISATIONS

A number of industries already feel the impact of mobile computing devices (Varshney, Vetter, & Kalakota, 2000). In universities, mobile technology infrastructures are being implemented so that students can access academic databases from any campus location (Willard, 2000). In government, police and criminal justice organisations use mobile computing technologies, as they need mobile access to information for law enforcement. In police organisations in the U.S., mobile computing terminals are used

for access to federal, state, and county records in order to facilitate auto registrations, summons, and warrants of arrest (Seaskate, 1997). These are also used for online offence reporting. In a recent study of the impact of these devices on the organisation, these mobile computing terminals enabled better communication among officers, increased the availability of information, and have been found to have a significant positive impact on officers’ job satisfaction (Agrawal, Rao, & Sanders, 2003). Gartner (Casonato, 2001) found in a survey of 212 respondents who had implemented mobile technology to support “business-to-employee” scenarios that the main benefits were increased employee productivity, followed by cost reduction and cost management, new information channel, and experimentation.

In the service industries that involve product delivery, for example, it is expected that mobile inventory management systems used to track the location of goods help improve delivery times and customer service. United Parcel Services has been an early adopter of wireless technology, using radio transmitter technology in trucks to send package delivery data back to the central UPS network, so customers can track package delivery in real time. It recently invested \$100 million to upgrade and consolidate its wireless network to Bluetooth technology in order to reduce operating costs. In a recent survey conducted among business-technology professionals, improved mobile technologies, business applications, and lower prices are factors that drive the use of mobile technology in business. The benefits cited by those surveyed were: increased employee communication and data sharing, increased employee productivity, improved customer service and satisfaction, easier collaboration with business partners, and increased access to corporate data for decision making (Ewalt, 2004).

IMPACT OF MOBILE COMPUTING

The benefits and impact of such technologies are factors that can be considered in the success of these mobile computing systems. In analysing the success of information systems, DeLone and McLean (1992) suggest that six dimensions of the system need to be considered: system quality, information quality, use, user satisfaction, individual impact, and organisational impact. These dimensions relate to each other in the following manner:

System Quality and Information Quality singularly and jointly affect both Use and User Satisfaction. Additionally, the amount of Use can affect the degree of User Satisfaction—positively or negatively—as well as the reverse being true. Use and User Satisfaction are direct antecedents of Individual Impact; and lastly, this Impact on individual performance should eventually have some Organisational Impact. (DeLone & McLean, 1992, pp. 83-87 as quoted in Myers, Kappelman, & Prybutok, 1998, p. 102)

This framework can be used in analysing the overall success of mobile computing systems. For example, in the case of Agrawal et al.'s (2003) study on the use of mobile computing systems in police organisations, it can be argued that the officers using the mobile technology gained mobile access to important law enforcement information. The infrastructure of the system provided a level of system quality and information quality as officers felt they were able to reliably communicate and share important information. With remote and mobile access, this promoted high use and resulted in satisfaction in using the system. The individual impact of this was felt in job satisfaction. Because of the speed of licence plate checks on the system, this helped deter criminal activ-

ity. Thus, it can be argued from this point of view that this mobile computing system made some organisational impact in that it helped in increasing organisational effectiveness.

Deloitte Touche (2001), in its management briefing on mobile technology, identified a barrier to the uptake of mobile technology being the limitations of the available mobile devices in that users often required access to data and voice simultaneously. The recent convergence of wireless technology has overcome this problem. Many single-purpose devices such as bar code readers, pagers, mobile communicators, and personal digital assistants (PDA) have now evolved and converged into single devices offering a broad range of functionality. This increased functionality and processing power has encouraged the development of various mobile application solutions. However, even with this improved functionality, the applications would be limited without the introduction of high-speed wireless networks. The increased coverage of the GSM/GPRS network has enabled companies to transfer large volumes of data in real time. Internally companies have the choice of wireless technologies such as 802.11b to interface to the local area network or Bluetooth to interface with other technologies. But the advent of improved communication technologies is presently geographically constrained, limiting the range of mobile applications while at the same time necessitating a range of communication technologies to be implemented in companies that operate in diverse geographic locations.

Even though many of the existing barriers have been addressed, there are still limitations associated with security and application development (Deloitte Touche, 2001). In terms of security, many companies are still coming to grips with issues within their companies rather than those caused by extending the boundary of the company onto developing technologies such

as mobile computing devices. Presently most mobile applications are merely extensions of existing enterprise applications, and therefore human computer interface issues pertinent to the mobile device are often overlooked.

IMPLEMENTATION CONSIDERATIONS

Companies have implemented mobile technologies to address a variety of business scenarios (ESRI, 2002; Symbol, 2002; Gedda, 2004). Deloitte Research (2002) developed a model which classifies the maturity or evolution of mobile applications into three different stages (see Figure 1). In the first stage, "Mobile Enablement," the existing applications interface is extended onto a mobile device. In this stage there is limited new functionality other than portability, but the device acts as an alternate input device. In the second stage, "Mobile Reinvention," probably due to the familiarity with the mobile solution, further efficiencies can be gained using a business process reengineering approach whereby new functionality can be achieved due to the nature of the mobile device. The final stage, "Mobile Discontinuity," is where new innovative busi-

ness scenarios can be developed which transform the organisation.

Up until now, very little research has been conducted in the area of mobile solutions, and their development and implementation in Australian companies. Even though many companies have or are in the process of implementing such solutions, the documentation of benefits and issues is lacking. This chapter attempts to document such a solution in an endeavour to provide a foundation for future case studies which can assist Australian industry with examples of the application of mobile technology.

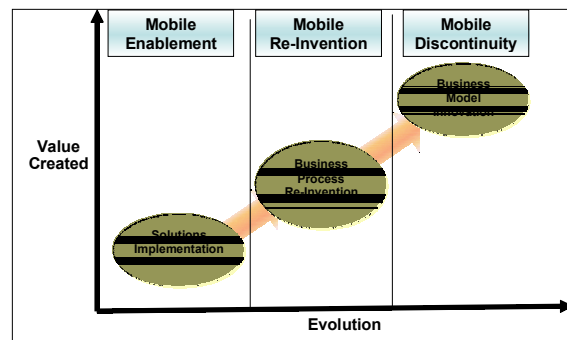
RESEARCH METHODOLOGY

This chapter adopts a case study approach to examine the impacts and issues associated with replacing a paper-based system with a mobile computing solution. Yin (1994, p. 35) emphasises the importance of asking "what" when analysing information systems. Yin goes further and emphasises the need to study contemporary phenomena within real-life contexts. Walsham (2000, p. 204) supports case study methodology and sees a need for a move away from traditional information systems research methods, such as surveys, toward more interpretative case studies, ethnographies, and action research projects. Several works (Chan & Roseman, 2001; Lee, 1989; Benbasat, Goldstein, & Mead, 1987) have used case studies in presenting information systems case-study research. Cavaye (1996) used case study research to analyse inter-organisational systems and the complexity of information systems.

The data-collection process for the present research included:

- examination of existing documentation,
- interview of actors, and
- direct observations.

Figure 1. Deloitte Touche mobile application maturity



CASE STUDY BACKGROUND

The Water Corporation has been in operation for more than 100 years, and is responsible for the provision of water and removal of wastewater, servicing more than 1.8 million people in the state of Western Australia. The corporation services both urban and rural communities spread across the 2.5 million square kilometres of the state. As part of this service, the corporation is responsible for the establishment and maintenance of the necessary infrastructure. This includes:

- 250 water treatment plants,
- 110 dams and reservoirs,
- 715 bores in 107 bore fields,
- 30,538 kilometres of water mains,
- 12,579 kilometres of sewers, and
- 2,782 kilometres of drains.

The corporation employs more than 2,000 staff, with the majority based in the state's capital. An important component of the information systems infrastructure utilised by the corporation is its ERP system. The corporation had been a user of SAP's R/3 ERP system since 1998 when they installed version 3.1H. In 2001 they upgraded to the latest version available (4.6c). The scope of this implementation included financial, human resource, plant maintenance, and project management modules, as well as SAP's data warehouse solution. The SAP system is used at 50 sites by more than 1,400 users, with 350 of these being concurrent at any one time.

A major responsibility of the corporation is the maintenance of the extensive range of infrastructure responsible for the delivery of water and removal of wastewater both in the rural and metropolitan areas. The workforce responsible for this is made up of employees from within the company and external contractors provided by a partner company. The exter-

nal contractors predominately worked within the metropolitan area. Traditionally, each day the workers would be allocated work orders for reported faults and new installations. This totalled more than 300,000 orders annually, with approximately 800 orders per day communicated via mobile telephones. The workers would collect the appropriate materials and maps for each job before travelling to the various work locations. Getting to the correct location was often difficult due to the vast network of pipes and the remote locations. The corporation had developed a geographic information system (GIS) that recorded the location of infrastructure items down the mains and manhole level of detail. Workers were provided with a printout of the necessary map for each work order.

Once field staff arrived at a location, they would assess the task and order any additional materials that may be required. This was done by telephone or radio. When the work was completed, employees were required to complete a range of paperwork. This included worker details, duration and type of job, materials used, and work completed. This paperwork was required for each allocated task.

On returning to their base office, the paperwork was collected and the data was entered into the SAP system. This was then used for necessary time calculations including payroll and job costing. The materials used were entered into the materials management module, and were used to assist with inventory of components and costing of work completed, which satisfied the plant maintenance requirements. The possible efficiencies from the capturing of this information were hindered due to the difficulties associated with capturing of this information in the field in a paper-based format. Due to the nature of the plumbing work and the less-than-suitable nature for recording information on paper in the field, often the paperwork was incomplete, illegible, or lost.

In 2001 it was decided to investigate how mobile computing devices could possibly overcome some of the issues associated with paper-based forms. A business case was developed with the objectives of:

- providing support for business objectives, and enhancing the efficiency and effectiveness of the key activities performed by field service delivery staff;
- enabling timely and resource-efficient scheduling of work;
- facilitating timely feedback to customers as to the status of work; and
- creating the capacity and flexibility for field support staff to meet the evolving business requirements of the corporation and its customers (Water Corporation, 2002).

The specific benefits identified included a reduction in data entry staff, reduction in errors resulting in an improvement in data integrity, improved scheduling resulting in a reduction of overtime hours, and reduction in field staff and vehicles.

It was expected that the project would enable efficient linkages, through mobile devices, to the corporate systems. The project would support 420 field crews. Of these, 150 service the Perth metropolitan region, while the remaining 270 service outlying rural regions. It was intended that the metropolitan employees would interact with the corporate systems in “real time” utilising GPRS technology, while the regional employees would synchronise with the corporate system in batch mode via the corporation’s wide area network or via dial-up connection. The overall project budget was expected to be \$AUD3.5 million, which included the typical implementation costs as well as costs associated with making staff redundant.

The main issues associated with the business case were application development, mobile device cost and functionality, infrastructure stability, transformation of business processes to mobile technology, and ongoing support. It was believed that due to the complexity of the solution, it was preferable to award the contract to a consortia of implementation partners who best addressed the above issues rather than awarding individual components to separate partners. This was believed to alleviate different partners, laying blame on other partners involved.

A request for proposal was distributed to interested parties and was eventually awarded to a consortia composed of Deloitte Consulting and Telispark in early 2003. They were expected to supply the mobile computing application licences and implement the mobile computing solution as well as providing ongoing support.

The Mobile Solution

The project commenced in April 2003 and was completed in December 2003. The solution enabled field staff to log on to their mobile computing device and display a list of work orders. They could then “*drill down*” to get more details of the order. The corporation’s PC-based GIS system had been further developed to operate in a mobile computing environment. Now instead of having paper-based maps for each work order, the solution enabled workers to enter a location into the device and then display the appropriate map. The maps included addresses, streets, suburbs, and infrastructure locations. The workers had the ability to zoom in or out depending on the amount of detail they required.

When they were about to leave for the site, they would click a button in the device to indicate that they had started travelling, and

once they arrived they clicked another button which then calculated travel time for each job. This also indicated that work had commenced. Once the job had been assessed, they would select the type of work from a predetermined list and materials required. When the job was completed or work had been halted for some other reason, this was also recorded in the device. The field worker would then display the next work order.

The information was synchronised between the mobile device and the SAP system using a variety of technologies depending on the type of mobile device and the communication technologies available in the area. In the metropolitan area, GPRS/GSM technologies were utilised which provided real-time synchronisation, while in outlying areas where communication technologies were limited the data was synchronised in batch mode using dial-up facilities when staff returned to their offices.

Technology

As mentioned previously a component of the project was the selection of appropriate mobile computing devices and the creation of the technological infrastructure to support the integration of these devices.

After a review of the available mobile computing devices suitable for the identified tasks, it was decided to implement two distinctly different devices. The first was a device made for rugged conditions (70%) which used the traditional telephone network (PSTN) for communication. The second was a sleeker and stylish device produced by Oxygen (XDA) which accounted for the remaining 30%. The XDA was a combination of a mobile phone with GPRS/GSM capabilities as well as the traditional PDA functionality.

In terms of the technological infrastructure, the complexity was considerable, with 21 com-

ponents from a variety of vendors. With so many interdependent components, an analysis was undertaken to identify single points of failure, whereby if a particular component failed, the solution would cease to operate. The analysis identified 12 separate components on which the solution was reliant, with the worst case scenario of a 70-hour recovery time for one particular component.

Implementation

From the commencement of the project, the importance of change management and training was identified. There needed to be a considerable change in culture to gain successful adoption of the solution. As part of the change management program, a short video was produced to explain the purpose of these devices and possible benefits they could provide. It also attempted to allay fears associated with increased scrutiny by management. Training courses were then conducted with staff at numerous locations, and they were then supplied with their mobile computing device.

One unexpected factor which contributed to the adoption of the devices was the selection of the XDA device. This device uses the Microsoft Pocket PC operating system which was supplied with a range of software including Microsoft Outlook used for contacts, appointments, and so forth, as well as entertainment software including games. When staff were issued their mobile devices, the GIS solution was implemented but the work order solution was not completed and therefore not implemented on the device. Initially it was decided to remove any unnecessary software other than those related to work tasks. But this did not happen and probably partly contributed to the success of the implementation. It encouraged workers to become competent with their device, as it could be used for a range of non-

work-related tasks. For many of the workers, the XDA had a novelty attraction, as very few of them had used such a device before and were impressed with its capabilities. The increased usage of the device facilitated the uptake of the work order solution when it was rolled out.

Interestingly the solution was more readily adopted by employees of the corporation compared to external contractors. It was felt that the contractors saw this solution as a means to better manage them and reduce their payments. Another unexpected factor is that the more “ruggedised” devices were not required, as minimal damage has been reported with the XDA devices.

In terms of business benefits at the time of the interviews, it was too early to assess. The adoption of the devices had been successful and preliminary results have indicated improved data integrity and plant maintenance reporting. Staff involved in managing the project have investigated other mobile solutions adopted by various utility companies throughout Australia; they believe that based on these examples, the corporation will achieve a good return on investment.

Already the corporation is considering extending the project to include further functionality. They expect to rollout PDAs in the future which incorporate global positioning systems (GPSs) to facilitate workers’ navigation to specific locations. An optional extra in the RFP document was automated vehicle location functionality. This would enable the corporation to identify which vehicle was nearest any particular fault, resulting in better response times and scheduling.

CONCLUSION

In terms of the mobile application maturity as per the Deloitte Touche model (Figure 1), the

Water Corporation’s mobile solution would fall into the “mobile enablement” category whereby an existing application or business process has been extended via mobile technology. The mobile solution is only in the initial stage of rollout and needs further investigation to assess the overall success and impact on the organisation. It is intended to revisit the organisation later in the year to conduct follow-up research. The issues that will be investigated will be the uptake of the various mobile devices, the robustness of the technical infrastructure, the tangible and intangible benefits, and future mobile solutions.

REFERENCES

- Agrawal, M., Rao, H. R., & Sanders, G. L. (2003). Impact of mobile computing terminals in police work. *Journal of Organisational Computing and Electronic Commerce*, 13(2), 73-89.
- Benbasat, I., Goldstein, D., & Mead, M. (1987). The case research strategy in studies of information systems. *MIS Quarterly*, 11(3), 215-218.
- Casonato, R. (2001). *Mobility and business to employee applications*. Retrieved May 2004 from <http://www3.gartner.com/pages/story.php.id.2545.s.8.jsp>
- Cavaye, A. (1996). Case study research: A multi-faceted approach for IS. *Information Systems Journal*, 6(3), 227-242.
- Chan, R., & Roseman, M. (2001, December 4-7). Integrating knowledge into process models—A case study. In G. Finnie, D. Cecez-Kecmanovic, & B. Lo (Eds.), *Proceedings of the 12th Australasian Conference on Information Systems*, Coffs Harbour, Australia (pp. 113-120).
- Dedo, A. (2004). *The return on your mobile investment*. Retrieved June 2004 from <http://>

- download.microsoft.com/download/1/a/5/1a572c42-10b5-469d-9acb-cedd2e634985/MobileDevices_ROI.doc
- Deloitte Touche. (2001). *Management briefing: Mobile technology*. Dublin: Deloitte Touche.
- Deloitte Research. (2002). *Mobilising the machine*. London: Deloitte Touche Tohmatsu.
- DeLone, W. H., & McLean, E. R. (1992). Information systems success: The quest for the dependent variable. *Information Systems Research*, 3(1), 60-95.
- ESRI. (2002). *Boulder County, Colorado, sign maintenance receives big dividends from ArcPad*. Retrieved June 2004 from <http://www.esri.com/news/arcnews/winter0102/articles/boulder-cnty.html>
- Ewalt, D. (2004, June 7). The pros and cons of wireless connectivity. *InformationWeek*. Retrieved from <http://www.informationweek.com/showArticle.jhtml;jsessionid=CCYMSQZTK4WR0QSNDBCSK0CJUMKJVN?articleID=21401672>
- Gedda, R. (2004, June). Mobiles make the grade. *Computerworld*. Retrieved from <http://www.computerworld.com.au/index.php/id;1903328574;relcomp;1>
- Lee, A. (1989). Case studies as natural experiments. *Human Relations*, 42(2), 117-137.
- Mobile Planet. (2004). *Mobile manufacturing: Capturing critical job tracking and inventory data*. Retrieved May 2004 from http://www.mobileplanet.com/askexperts/solutions/syware_case10.asp
- Myers, B. L., Kappelman, L. A., & Prybutok, V. R. (1998). A comprehensive model for assessing the quality and productivity of the information systems function: Toward a theory of information systems assessment. In E. J. Garrity, & G. L. Sanders (Eds.), *Information systems success measurement*. Hershey, PA: Idea Group Publishing.
- Nelson, M. G. (2001). Wireless delivers for UPS overhaul. *Informationweek*, 109-113.
- Paavilainen, J. (2001). *Mobile business strategies: Understanding the technologies and opportunities*. London: Wireless Press.
- Seaskate. (1997). *The evolution and development of police technology*. National Institute of Justice. Retrieved from <http://www.nlectc.org/virlib/InfoDetail.asp?intInfoID=169>
- Strategis Group. (2001). *Mobile data penetration to be nearly 60% by 2007 says the Strategis Group*. Purchase, NY: Strategis Group.
- Varshney, U., Vetter, R. V., & Kalakota, R. (2000). Mobile commerce: A new frontier. *IEEE Computer*, 32-38.
- Walsham, G. (2000, June 9-11). Globalisation and IT: Agenda for research. In R. Baskerville, J. Stage, & J. DeGross (Eds.), *Organisational and social perspectives on information technology*, Aalborg, Denmark (pp. 195-210). Boston: Kluwer Academic.
- Water Corporation. (2002). *Request for proposal document for the provision and implementation of a mobile computing system*. Perth, Western Australia: Water Corporation.
- Water Corporation. (2003). *Annual report*. Retrieved from <http://www.watercorporation.com.au/annrep/index.cfm>
- Willard, C. (2000). High wireless act. *Computerworld*, 34(31), 49.
- Yin, R. (1994). *Case study research, design and methods* (2nd ed.). Newbury Park, CA: Sage.

Chapter LXI

Introducing Mobile Technology into an Australian City Council: Experiences and Lessons Learned

Joanne Marie Curry

University of Western Sydney, Australia

ABSTRACT

In an ongoing bid to provide high quality local government services, Penrith City Council partnered with the University of Western Sydney to derive a mobile strategy for the development of a range of handheld systems for use in the field. Several R&D projects aimed at determining the viability of using mobile technology for the conduct of off-site health, building and development and sewerage inspections and the allocation of parking and waste management infringements were conducted over a two-year period. Some significant issues relating specifically to the implementation of mobile technologies in a large Australian city council were encountered including: release hype vs. the implementation realities of mobile technology, technological options for the introduction of mobility, user acceptance of new technologies, management of client expectations, and local government standards and guidelines and their impact on development directions. The experiences and lessons that were learned from these projects can be of assistance to other local government agencies and similar organisations employing a heterogeneous workforce that is restrained by external legislation and policy.

INTRODUCTION

Penrith is a city on the western fringe of the Sydney metropolitan area, in east central New South Wales, Australia. Located at the foot of the Blue Mountains just 55 kilometres from the Sydney CBD, the City of Penrith covers an

area of some 407 square kilometres. With an estimated resident population of approximately 178,000 (as at June 2004), there has been a 15% growth rate since 1991 (*Australian Bureau of Statistics*, 2004). With such rapid expansion and approximately 4,500 businesses operating in the local government area, Penrith

is a fast-growing metropolitan region, and in just a generation has grown from a rural town to the major regional service centre for the outer western Sydney area. In the bid to continue to provide high-quality local government services, the IT Department of Penrith City Council, working in partnership with the University of Western Sydney, undertook several projects aimed at determining the viability of using mobile technology for Council operations such as the conduct of off-site health, building and development, and sewerage inspections, and the allocation of parking and waste management infringements.

The introduction of the Penrith City Council Mobile Strategy evolved over a two-year period and provided some interesting experiences for the development teams and contributed significantly to their learning in many areas, including: release hype vs. the implementation realities of mobile technology, technological options for the introduction of mobility, user acceptance of new technologies, management of client expectations, and local government standards and guidelines and their impact on development directions (Bryan, Holdsworth, Sharply, Curry, & McGregor, 2002; Curry & Lan, 2004). This chapter provides highlights of the development work performed by “high-achieving” student groups, and discusses some of the resultant experiences and the detailed lessons learned from this work of incorporating mobility within the Council’s business processes.

PROJECT BACKGROUND

In the mid-1990s Penrith City Council (PCC) entered into a partnership with the University of Western Sydney (UWS) for engaging final-year computing students for Research and Development and Prototyping projects. This chapter concentrates and expands on a few select “high-achieving” student projects deal-

ing with incorporation of mobility within the Council’s business processes that were conducted between March 2002 and December 2004. These are invaluable projects for both their academic importance as well as their practical importance to the Council, wherein their overall aim is to reengineer existing processes for PCC field officers required to complete official documentation off-site. Field officer work ranged from inspections for development and building applications, food and hygiene surveillance reports, and issuing of parking and general counsel by-law infringement notices.

Penrith City Council is required to make many inspections in areas such as health, building and development, and sewerage. The management of the data collection and data entry for inspections is becoming increasingly cluttered and time consuming using the existing manual system.

The current system requires a field officer (inspector) to attend an inspection site equipped with the necessary paper forms for collecting data. These forms are completed in duplicate on site: one copy for the subject of the form (e.g., the owner of the premises being inspected), and a duplicate copy for Council’s records. Once all inspections have been completed for the day, the inspector returns to the office where the data for each inspection is typed manually into the database, either by the field officer or an administration clerk.

This resulted in double handling of the data, thus increasing the general workload for Council administrative staff, as well as making the duplicate form redundant once the data was entered into the PCC database. Data entry errors were also common, leading to corrupt data residing on the Council database leading to poor quality management reporting.

One of the key PCC goals in the achievement of this reengineering initiative was the introduction of mobile technology to the new

process environment. The justification for this approach, based on Gronlund (2002) and Garson (1999), can be given as follows:

- **Improved Security of Documentation Completed Off-Site:** Paperwork would not be misplaced as it would be securely stored in the mobile device.
- **Easier Off-Site Information Retrieval:** Information could be easily retrieved and updated if errors were identified while the field officer was still off-site.
- **Reduced Volumes of Manual Data Entry:** Automatic download of off-site data will dramatically reduce the time need for manual data entry. PCC had estimated the tangible benefits for reduced data entry at over \$15,000.
- **Reduced Data Entry Errors:** Results of off-site work could be transferred to the PCC database automatically, thus significantly reducing the amount of incorrect data held.
- **Reduction of Physical Paper Volumes Requiring Storage:** Hard copies of reports are only printed as needed.
- **Improved Management Reporting:** Due to reduced data entry errors, the quality of resulting management reports would increase.

Several general environmental constraints related to the processes used by the Council's field officers were identified including:

- Field officers are required to be extremely mobile and therefore cannot be burdened by heavy and/or large equipment.
- Field officers are out in the elements everyday, therefore the solution must be unaffected by rain, heat, and other normal working conditions an officer experiences.

- There may or may not be access to a continuous power source, therefore the solution must be self-dependant for a substantial length of time.
- Due to the nature of the business (government agency), the solution must also be very secure with regards to data.

In order to address the Council's problems, the system needed to include the following capabilities:

- Ability to download lists of inspection appointments to be completed by the field officers, from the Council database to a handheld device.
- Capture an electronic version of data currently being recorded on paper forms.
- Automatically store captured electronic data into the relevant fields of the applicable Council database.
- Allow for a paper copy of the data to be produced for issue to the client upon request.

During the period March 2002 to December 2004, six selected student project teams worked on six phases of the PCC Mobile Strategy (see references for the project reports from these teams). Each team was required to follow a formal Systems Development Lifecycle, including production of system documentation for ongoing system development and maintenance by PCC IT staff (Curry & Stanford, 2005). The remainder of this chapter will detail the work completed in each of the six phases, and will highlight experiences and lessons that were learned that are relevant to the introduction of mobile technology into a local government environment.

PHASE 1: FOOD AND HYGIENE SURVEILLANCE REPORTING VIA A REMOTE INTEGRATED SYSTEMS TERMINAL (MARCH 2002)

The scope for Phase 1 was to develop a mobile systems architecture that catered to the development of an automated data entry system to be used for building and development inspections and food and sewerage inspections, using handheld devices. The first module to be developed dealt with inspections for food production premises and collected data relating to food production and hygiene standards.

The new system was to be designed to suit the unique needs of Penrith City Council. This involved data being input directly to a PDA, with the resulting files being transferred across the palm top operating system to the Council system, with the information being downloaded directly into the Council database. It was proposed that for the first-stage prototype, Universal Serial Bus (USB) connections to upload the data from the handheld device would be used. Bluetooth or similar technology would be researched at a later date to enable the uploads to be performed remotely. The initial application used Microsoft's eMbedded Visual Basic 3.0 as the development platform with a Microsoft Access database. ActiveSync was used to manage the USB connection and synchronise the data transfers.

The Remote Integrated Systems Terminal (R.I.S.T.) handheld system was designed on the Compaq IPAQ 3870 using Microsoft's Pocket PC version 3.0.11171 (Build 11178). The device ran the Arm SA1110 CPU, as well as being Bluetooth enabled (for the future).

The primary purpose of the first user interfaces was to make them look like the original paper-based form the field officers were used to. Only minor cosmetic changes were made where necessary.

Project Progress

The initial phase saw the development of the underlying architecture for the use of handheld devices for the PCC field officers with comprehensive development of the Food and Hygiene Surveillance Report application. The R.I.S.T. application was developed to a Beta software level, with the high-priority modules readied for implementation. These included the recording of the inspection data, and the synchronisation between the database on the Pocket PC and the database on the host PC (Team S02106A, 2002).

PHASE 2: ENHANCEMENT TO PHASE 1 AND DEVELOPMENT OF THE SEWERAGE INSPECTION MODULE (AUGUST 2002)

The following semester the R.I.S.T. project entered its second phase. The R.I.S.T. beta software application (Phase 1) had been deemed successful by the IT Manager and his staff as an R&D activity, and they were now prepared to enhance Phase 1 and develop another "inspection" module ready for release to the field officers.

The second phase of the development focused on turning the beta software into a commercial quality application, and adding extra functionality to the existing application. Once this was complete, the team was to begin work on the next inspection module—sewerage.

Initially the existing MS Access database needed to be reviewed due to some intermittent data entry errors, which were resulting in the R.I.S.T. application not being able to synchronise with the host PC for every upload. A new transition table was created that was only accessed by the Pocket PC. The new table was then used in the data upload to the PCC database.

The initial version of the Food and Hygiene Surveillance Report application was only of beta quality. This meant that the software contained the basic functions that it required but still needed some finetuning to make it of implementable quality. Testing of the beta software had yet to start in the field, and it was envisaged that further modifications might be needed if the field officers were not completely satisfied.

Slight modifications were made to the beta software, with extra functionality being added to the interface. Modifications to the report formats were also made.

Once the Phase 1 enhancements outlined above were complete, development work began on the second inspection module: Sewerage Inspection. Development work followed the initial architecture and basic functionality of Phase 1. Field officer testing began and an issue was identified with re-inspection visits, which had not been previously catered for. As re-inspections dealt with an existing inspection report and only a minority of items needed updating, all inspection details from previous inspections needed to be retrieved to the PDA prior to the field officer attending the inspection site. A new database table was added to the existing database, and the data from four other tables were integrated into the new table. Following the inspection, updated data from the new table was segmented as needed and updated at the original source.

Project Progress

The Sewerage Inspection module was completed and thorough testing by the inspectors commenced. Some minor bugs were reported and, once these were overcome, PCC management directed that the new PDA-based inspection system be implemented in the field (Team P02203A, 2002).

PHASE 3: BUILDING INSPECTION MODULE DEVELOPMENT (MARCH 2003)

Following the success of Phases 1 and 2, PCC decided to expand the new mobile architecture into the larger and more complex area of building and development applications (BA/DA). PCC already had an advanced online submission system for building and development applications, and was looking to further its technological leadership by enabling the building inspectors to carry out mobile inspections. With the increased growth in industry and population in the Penrith Valley area, the number of building and development application inspections had amplified significantly and now averaged 30-40 per day. This had brought with it similar problems for the manual system as outlined above. Most significantly it had been estimated by PCC that the time that could be saved by automating the transfer of building inspection data to the Council database could save in excess of \$15,000.

The purpose of Phase 3 was to create a mobile system that leveraged the existing systems and equipment into which Penrith City Council had already invested time and money. Development progressed smoothly and the solution was handed over for field-testing as scheduled.

Project Progress

During early testing of the Building Inspection module by the building inspectors, the inspectors expressed some initial concerns about the technical complexity of the solution, the small screen size of the PDA, and the resulting difficulty they had in reading and completing the online forms. The building inspectors expressed a view that they would be much happier completing a hardcopy form. They asked if

Table 1. System architecture at the completion of Phase 3

<i>Hardware</i>	<i>Software</i>	<i>Development Environment</i>
• Pocket PC iPAQ 3900 • Windows Desktop PC	• ActiveSync 3 • MS Access 2000	• Pocket PC 2000 Software Development Kit • eMbedded Visual Tool 3.0 ○ eMbedded Visual Basic ○ eMbedded Visual C++

there was some way that this form could be transmitted to the PDA automatically and then uploaded to the Council database as originally planned. This required the project team to complete some additional research into other technical options that could accommodate these needs and that were more acceptable to the current user group (Team S03101A, 2003).

The system architecture at the completion of Phase 3 is shown in Table 1.

PHASE 4: PARKING INFRINGEMENT MANAGEMENT MODULE

Phase 4 of the PCC Mobile Strategy involved development of a Parking Infringement module for the handheld device. The added complexity for this module revolved around the fact that a hard copy of the parking infringement was required to be left on the windshield of the offending vehicle.

Initial investigations had found that only one other council in Australia was currently using PDA technology for the issue of parking infringements, but that the system was still in its infancy and not ready for commercialisation.

Following significant research, the development team recommended a solution that used the existing iPAQ with a Thermal Micro Printer. The printer unit from Unique Micro Design's PP-50 range was chosen because of its slimline and ergonomic design. The Datec PP-50 is a portable cradle type printer for a variety of palm and pocket PC handheld-type computers (Unique Micro Designs, 2005).

The unit can also act as a Hot Sync cradle and a recharge station for rechargeable devices. Software utilities, Printer Manager, Font Manager, Screen Pad, Logo Manager, and Print Emulator were also included, and Software Developers Kits (SDKs) were available for MS-embedded VD, C/C++, and .NET Compact Framework.

The unit features included: ergonomic handheld design; fast, direct thermal printing; data, barcodes, and graphics printing; long-life rechargeable battery; and 32- or 42-column printing.

Project Progress

The resulting solution delivered the new Parking Infringement module with the added facility of printing a copy of the parking infringement for attachment to the offending vehicle. Additional features included a desktop content management system for the management of parking officer staff details and infringement types.

Figure 1. Example of a Thermal Micro Printer with iPAQ inserted (paper printout also shown)



The solution was considered functional, light-weight, compact, and relatively user friendly. It also provided PCC with a strong business potential for on-selling the resulting system to other interested parties such as councils, national parks and wildlife bodies, universities, shopping centre car parks, and any other major public venue where parking restrictions existed. The project team highlighted that the mobile parking infringement market was virtually untouched in Australia, and any available systems were based on overseas designs and therefore were not suited to Australian regulations and legal practice. PCC accepted this recommendation under advisement and proceeded to investigate the full commercialisation of the newly developed system (Team S03204A, 2003).

PHASE 5: NEW WASTE MANAGEMENT INFRINGEMENT MODULE, AN INTEGRATED GENERAL INFRINGEMENT MANAGEMENT SYSTEM AND REMOTE UPDATING RESEARCH (MARCH 2004)

The initial work in Phase 5 leveraged on the existing PDA development work and saw the

construction of a new Waste Management module. This new module was then integrated with the existing Parking Infringement module to create a General Infringement Management System for the creation and management of all off-site infringement notices issued by PCC field officers. Storage of field officer signatures and the ability to include the issuing officer's signature on an infringement notice was also incorporated.

The second stage of this phase involved research into a remote updating capability, with the aim of further reducing the need for the field officers to personally attend the Council offices. As part of this research, PCC also asked the student team to look for options that could overcome the issue of not being able to leave a hardcopy of the completed building inspection form with the client onsite.

Project Progress

The project team successfully implemented the new Waste Management module and integrated it with the existing Parking Infringement module to create the PCC General Infringement Management System for handheld devices. Research into the remote updating capability and production of an onsite hardcopy of the completed building inspection form led to the

Figure 2. Example of User Interface Design for Parking Infringement module

The figure displays two screenshots of the Parking Infringement module's user interface. The left screenshot, titled 'Record Infringement', shows a form for entering new infringement data. It includes a 'Location Details' section with fields for 'Time of Offence' (radio buttons for 'At:' and 'Between:'), 'Date' (with a 'Today' checkbox), 'Carpark Name', and 'Suburb'. Below this is a 'Notice status' dropdown menu and a 'parking from' field. A 'Menu' button is at the bottom. The right screenshot, titled 'Infringement Summary', displays the details of a specific infringement. It includes fields for 'Vehicle' (987-ACF (QLD) Holden Vectra, Green), 'Date and Time' (Fri 3 Oct 2003 Between 11:50 am and 12:30 pm), 'Location' (High St b/n Riley and Station St, Penrith), 'Offence Code' (6660), and 'Penalty' (\$143.00). The summary also includes the text 'Stand vehicle in disabled persons parking space without authority' and buttons for 'Print Infringement', '< Back', and 'Mark for Cancel'.

presentation of three alternatives and the recommendation that the Council look at the purchase of SmartPad devices combined with new GPRS-enabled PDAs (Team S04104A, 2004).

PHASE 6: REENGINEERING OF THE BUILDING APPLICATION INSPECTION MODULE AND IMPLEMENTATION OF REMOTE UPDATING (JULY 2004)

During testing of the initial PDA-based Building Inspection module, the building inspectors had expressed a certain degree of dissatisfaction. Functionally they felt that it was vital that a copy of the inspection result be left onsite with the client, and usability wise they were uncomfortable with the PDA technology and the size of the information they were required to work with (see above, and “Experiences and Lessons Learned below). This necessitated reengineering of the existing application to better accommodate these new requirements. PCC also took this opportunity to enhance the Building Application module to include the remote updating facility researched in the previous phase.

After reviewing the research outcomes of Phase 5 and confirming whether any other options provided further advantages, the team suggested that PCC take a different technological direction and utilise the facilities offered by a SmartPad system with a GPRS-enabled PDA (Seiko Instruments, 2005). Although this appeared to be a backward step technologically, it provided a simple solution to both of the highlighted issues.

The new process required the building inspectors to record the inspection results through the SmartPad interface. This was done by placing the SmartPad’s digitising tablet sheet under the Inspection results form and clipping the form directly onto the SmartPad’s tablet. The inspectors then completed the form with an

ink pen, just as they normally would during the course of the building inspection. Once the inspection is complete, the inspector removes both the inspection results form and the digitising tablet sheet—leaving the inspection results onsite, thus removing the need to print an extra copy for the client onsite.

The inspector then removes the digitising tablet, and changes the mode of the pen to stylus—which enables him/her to write directly onto the SmartPad. The inspector saves the recorded results and sends them (via infrared) to the accompanying PDA.

The formatting and transmission of the data to the PDA and the subsequent transmission of data to the PCC database did raise some interesting technical issues however. At the time of this project, the SmartPad was limited to saving the written results from the tablet into Graphics Interchange Format (GIF). The PDA in conjunction with handwriting recognition software then translated the .gif file, and prepared it for transmission. When ready, the inspector sent the text-based results back to Penrith City Council via GPRS. Once the results were received, OCR (Optical Character Recognition) software converted them into a readable format and saved the results into the database.

More specifically this involved:

- The results actually being e-mailed via GPRS in a .bmp format to a dedicated e-mail address at PCC’s office.
- A new module R.I.S.T. 2 was installed on the Council’s PC. This is a timer-based software that performs hourly polling, looking for any new building inspection results received in the inbox.
- R.I.S.T. 2 then extracts the .bmp files from the Microsoft Outlook inbox into a specified directory.
- The .bmp-formatted files are then converted into .tiff format.
- The iReadForms software is activated.

Figure 3. Example of SmartPad and GPRS-enabled PDA environment



- iReadForms then translates the results and creates a .txt file.
- This file is then transformed into an MS Access readable format and creates a temporary database table.
- The results from the temporary table are then uploaded to the PCC database.

Once the R.I.S.T. 2 system had completed the data transfer and manipulation, the information is then ready for processing as per normal Council guidelines.

Figure 3 shows an example of the mobile office environment offered by the SmartPad solution.

Project Progress

The requirements were fulfilled (not without some angst on the student's side!), and the new system approach was accepted by both the IT and building inspection staff (Team S04208A, 2004).

The resulting architecture now included:

- Active Sync Version 3.5
- Microsoft .NET Framework 1.1
- iReadForms Evaluation 4.1 Build 166
- iMate Personal Digital Assistant (PDA) with In-Built GPRS Capabilities

- InkNote Manager Version 3.2
- R.I.S.T Version 2.0
- Seiko SmartPad2 (Model SP582)

EXPERIENCES AND LESSONS LEARNED

Over the two years of these development projects, Penrith City Council, along with the students and UWS academics, all learned a great deal about the nuances of implementing mobility into a local government environment. Specific areas of this process of implementing mobility include: release hype vs. the implementation realities of mobile technology, technological options for the introduction of mobility, user acceptance of new technologies, management of client expectations, and local government standards and guidelines and their impact on development directions.

Release Hype vs. the Implementation Realities of Mobile Technology

The students were very excited to be working with mobile technology because it was new and leading edge, and they were keen to present the client with options that were at the forefront of the mobility movement. As with most new technologies, however, there is a time lag in the software and hardware features publicised by the vendors in their press releases and the implementation support available for these new features. The students, and their client, learned that trying to be on the "bleeding edge" of a new technology brings with it specific development issues and a certain degree of rework. This is evidenced in Phase 6 and the lack of support for form-based input. This point also impacted the development client's expectations (see below).

Technological Options for the Introduction of Mobility

In conjunction with the point above, there were always several options available for implementing the different features of mobility required by PCC. As evidenced in the introduction of the SmartPad option, what appeared to be an ideal solution on the surface required quite a number of “work-arounds” to actually achieve the client requirements.

User Acceptance of New Technologies

Although IT management at PCC were very supportive of the mobile strategy, they did not fully investigate the actual users’ acceptance of mobile technology into their working environments. The primary driver in the reaction of the different user groups was “age”. The field officers involved in issuing parking infringements were quite a bit younger than the building inspectors and had no problems using the PDA with the thermal printer for their duties. The building inspectors were quite a deal older, and their reactions were not as positive due to the size of the font they were required to work with and the perception that the new technology was too complex. As with any technology, it is vital to assess the users’ acceptance of the new environment and develop specific training to overcome this, or as in PCC’s case, realign the development direction to accommodate the users’ concerns.

Management of Client Expectations

The IT staff at PCC were very active in identifying potential new technology that would give them a competitive advantage and deliver their clients (the residents of the Penrith Valley area) the highest quality services. At times this led to excessive client expectations of what

was possible, and the students and UWS academic staff spent a good deal of time discussing how some of these expectations would impact the delivery schedule if they were pursued. The students also learned that all requirements identified as high risk (new technological area) needed to be fully investigated before development commitment was given to the client.

Local Government Standards and Guidelines and Their Impact on Development Directions

For most of the students, this was their first ‘real-world’ development experience. In addition to the normal learning curve for such situations, they were also faced with having to accommodate specific conditions that were imposed on them in relation to local government policy and bylaws. This affected the development work in several ways. In most cases the teams were not permitted to change the input forms that the field officers completed, as the Local Government Association defined these and they were standard across the state. This affected the user interface design on the PDAs on many occasions. The teams learned that in certain situations, external factors that were out of their control significantly impacted some of the potential quality improvements that would have been possible.

CONCLUSION AND FUTURE DIRECTION

The work undertaken between Penrith City Council and the University of Western Sydney was very rewarding for both parties. PCC was able to advance its mobility R&D efforts significantly by using low-cost resources (students). The Council was able to implement its mobile strategy considerably earlier than originally planned, thus delivering advanced high-

quality services to local communities. The students from UWS participated in projects at the leading edge of the mobility movement and gained invaluable experience for their future careers. Academic staff were also able to integrate the experiences and lessons learned into their teaching material for other students. The mobile strategy work is continuing, with PCC currently investigating the possibilities for commercialising the Parking Infringement module for on-selling to other councils and venues with existing parking restrictions.

REFERENCES

- Australian Bureau of Statistics. (2004). *Census data*. Retrieved April 20, 2005, from www.abs.gov.au
- Bryan, G. M., Holdsworth, D., Sharply, R., Curry, J., & McGregor, C. (2002, January 7-10). Using XML to facilitate information management across multiple local government agencies. In *Proceedings of the 35th Hawaii International Conference on Systems Sciences (HICSS'35)*, Big Island, HI (pp. 119-128).
- Curry, J. M., & Lan, Y. (2004, May 23-26). Web-enabled business operations: Development of the Semantic Web evolution model. In *Proceedings of the 15th Information Resources Management Association International Conference (IRMA2004)*, New Orleans, LA (pp. 1123-1125).
- Curry, J. M., & Stanford, P. (2005). *Practical systems development—A project-based approach*. Sydney: Pearson Education Australia.
- Garson, G. D. (1999). *Information technology and computer applications in public administration*. Hershey, PA: Idea Group Inc.
- Gronlund, A. (2002). *Electronic government, design, applications and management*. Hershey, PA: Idea Group Inc.
- Seiko Instruments. (2005). *Product description, SmartPad2*. Retrieved April 2, 2005, from <http://www.siibusinessproducts.com/products/sp582.html>
- Team P02203A. (2002). *System Development Lifecycle documentation, Computing Projects Archive, UWS Library*. Retrieved from <http://library.uws.edu.au/>
- Team S02106A. (2002). *System Development Lifecycle documentation, Computing Projects Archive, UWS Library*. Retrieved from <http://library.uws.edu.au/>
- Team S03101A. (2003). *System Development Lifecycle documentation, Computing Projects Archive, UWS Library*. Retrieved from <http://library.uws.edu.au/>
- Team S03204A. (2003). *System Development Lifecycle documentation, Computing Projects Archive, UWS Library*. Retrieved from <http://library.uws.edu.au/>
- Team S04104A. (2004). *System Development Lifecycle documentation, Computing Projects Archive, UWS Library*. Retrieved from <http://library.uws.edu.au/>
- Team S04208A. (2004). *System Development Lifecycle documentation, Computing Projects Archive, UWS Library*. Retrieved from <http://library.uws.edu.au/>
- Unique Micro Designs. (2005). *Product description, Thermal Microprinter*. Retrieved April 5, 2005, from http://www.umd.com.au/itd/products/datecs_pp50series.html

Chapter LXII

Emerging Mobile Technology and Supply Chain Integration: Using RFID to Streamline the Integrated Supply Chain

Richard Schilhavy

University of North Carolina at Greensboro, USA

A. F. Salam

University of North Carolina at Greensboro, USA

ABSTRACT

This chapter explores how a mobile tracking technology is able to further streamline the integrated supply chain. Previous technologies which have attempted to integrate suppliers, manufactures, distributors and retailers have lacked the flexibility and efficiency necessary to justify the prohibiting costs. Radio frequency identification (RFID) technology however enables various organizations along the supply chain to share information regarding specific products and easily remotely manage internal inventory levels. These applications are only a sample of what RFID is able to accomplish for the integrated supply chain, and this chapter seeks to explore those applications.

INTRODUCTION

This chapter sets forth to provide a holistic view of how a recently adopted wireless identification technology, specifically radio frequency identification (RFID) tags, could potentially revolutionize the integrated supply chain. Companies are able to become more flexible and

efficient by using a combination of mobile technologies and RFID to provide for remote inventory control and real-time, information-rich tracking of shipments in the distribution channel (Lapide, 2004). Although this technology has several hindrances currently blocking it from mass usage (Thompson, 2003), recent advancements in the technology have increased the

viability of RFID for widespread organizational use, increasing the capacity and strength while decreasing the size and cost. RFID now rests in a unique position wherein large organizations are strongly considering its viability in a variety of applications to streamline the supply chain.

Organizations have already begun considering its application in the realm of supply chain management, attempting to further streamline the process. However, while many authors have discussed the benefits of RFID tags for parts of the supply chain, this insight has only focused on a localized level, such as inventory management in retail outlets (Atkinson, 2004; Lapide, 2004; Kinsella, 2003; Schindler, 2003). Much of this discussion is centered on reducing costs for those isolated parts of the supply chain. For example, several large manufacturers are pushing the technology by actively conducting trials in manufacturing, distribution, and even retail. These companies include Proctor and Gamble, Gillette, Unilever, and retail giant, Wal-Mart (Kinsella, 2003). These RFID trials have been limited to single stages of the supply chain, focusing on the reduction of costs as the ultimate goal. Although cost reduction is commendable, true improvements in value for industry and consumers come through a unified effort to improve the entire supply chain network, reducing costs and improving accuracy and efficiency for all companies integrated into the network.

This chapter will first provide an overview of the technical aspects of RFID. Following this, an analysis of two perspectives of the integrated supply chain will be framed in light of the current and possible future applications of RFID in each area and the relationships between those areas. Finally, RFID will be framed in a holistic view of both the integrated supply chain as well as the demand chain, addressing some inter-organizational issues.

BACKGROUND

The RFID Tag and Reader

The core of RFID technology consists of two components, the identification tag and the tag reader. The identification tag itself is composed of a small antenna and a microchip, which stores a small amount of information pertinent to the object tagged (Rappold, 2003). Although the information stored may take a wide variety of forms, for many objects a simple code would be sufficient to identify the item. Asif and Mandviwalla (2005) identified five types of RFID tags in their RFID Applications Framework, including active, semi-passive, passive, chipless, and sensor. Tag readers may also be stationary or mobile, depending on the application. Of those tags which contain chips, RFID tags may either be active, passive, or a combination of the two. Active tags are powered by some external power source, such as a small battery. Passive tags, on the other hand, have no individual power source and receive power from the electromagnetic waves the tag reader uses to access the information from the tag. Some tags may use a combination of the two strategies, where an active tag containing a battery is recharged by the transmission used to read it. Chipless tags have the lowest power consumption, range, as well as cost of all the types of RFID tags since they do not contain either a battery or a silicon chip. Information storage is also significantly less, often only enough to store a simple product code.

The tag reader uses electromagnetic waves in the radio frequency band to transmit the data stored on the identification tags to the reader and, in some cases, power the identification tags. The reader may be a mobile or stationary unit depending on the application, and an organization could easily employ both. Mobile readers naturally benefit from being able to change

location; however, power limitations have a severe impact on the range and may even become an issue if passive RFID tags are used extensively. The effective range of a tag reader is a function of the frequency the tag reader is operating on and the power output available. At lower frequencies, range is severely diminished; however, power output is minimal. At higher frequencies, identification tags are able to be accessed from further away, but require significantly more power. Here, active RFID tags may need to be utilized to increase the range along with increasing the power output of the tag reader. Finally, sensor tags combine a small sensor targeted at a particular purpose, such as measuring temperature, viscosity, movement, and so forth, with an antenna, chip, and battery to store and transmit information from the sensor to a reader or network.

Not unlike any other wireless technology, RFID comes with a few limitations or issues affecting communication reliability (Angeles, 2005). The use of radio frequencies becomes a significant issue since these bands are often open to a multitude of other devices, such as wireless phones, computers, radios, and other office equipment. These common workplace devices may cause interference with reading a RFID tag and should not be overlooked when problems arise. Another problem common among mobile technologies is the increased collision of packets when more senders and receivers (in this case, more tags) are present. Since the reader is not limited to line of sight, a reader may pick up a multitude of tags in any direction of the reader, and if a large amount is present surrounding the reader, the transmissions between identification tag and reader may become interrupted due to such collisions. Anti-collision technologies are currently being developed to confront this problem with collisions (Angeles, 2005).

The RFID Network

Thanks to the Auto-ID Center at MIT, recent developments in RFID technologies have expanded RFID technology to create a holistic product identification system that consists of four components (Rappold, 2003; Smith & Konsynski, 2003; Asif & Mandviwalla, 2005). The identification tag and tag reader, again, transmit and read the information stored on the identification tag attached to the particular physical object. Stored on the identification tag is an electronic product code (EPC), which identifies the particular object or the state of that object. The EPC is a 96-bit identification code similar to conventional bar codes which uniquely identifies a product. The object name server (ONS) is a local or remote server that acts as a directory service, mapping the EPC to additional information about the physical object. This additional information in Physical Markup Language (PML) provides a standard format for describing products and storing other information about them (Smith & Konsynski, 2003). For example, the PML documents may contain information about the product manufacturer, source, and destination, or simply more detailed information about the product itself (Rappold, 2003). By mapping the EPC code to the PML documents containing information about the product, the tag may be significantly smaller since all the information is not required to be stored on the tag itself.

RFID Applications

Companies are able to become more flexible and efficient by using a combination of mobile technologies and RFID to provide for remote inventory control and real-time, information-rich tracking of shipments in the distribution channel (Lapide, 2004). Although this technology has several hindrances currently blocking it

from mass usage (Thompson, 2003), the potential long-term benefits are astounding for both the integrated supply chain and other mobile technology applications as well. RFID technology is able to provide item- and product-specific information which remains with the physical object. Since no line-of-sight is required to read a tag and multiple tags may be read simultaneously, inventories may be tabulated quickly with little manual labor and items may be tracked regardless of their location in the range of the readers. Rich information can be stored on the tags themselves, or simply mapped to the tags via an EPC, allowing this new technology to be easily mapped into current systems. PML provides additional information through a standardized markup language, providing additional interoperability between existing systems and the systems of other organizations. The technology is small, flexible, and relatively inexpensive. In the following sections, we will look more closely at the applications of RFID—specifically in the supply chain—and analyze the relative costs and benefits for each.

RFID AND THE SUPPLY CHAIN

One of the problems of current implementations of RFID tags and readers in the supply chain is that they have largely been efforts of a single company operating independently in their area of the supply chain. Technology improvements in the supply chain which are isolated to a single stage, such as manufacturing or retail, are limited to minor improvements in costs. To further illustrate this point, the following paragraphs will explain potential implementations of RFID in each stage of the supply chain considered entirely in isolation. Therefore, suppliers will be considered apart from manufacturers, retailers apart from distributors and consumers, and so on. Relative costs and benefits will be

weighed with each implementation, as well as the possible risks and rewards in undertaking the endeavor.

Suppliers

A common theme among many of the stages of the supply chain is inventory management, even when considered in isolation. Technology improvements in inventory management allow for significant improvements in labor and capital costs, and accuracy over traditional inventory systems which require manual operation. Suppliers are no different, requiring the maintenance of large amounts of raw and processed materials in various forms. However, there are unique considerations for each stage of the supply chain in regards to inventory management which needs to be addressed. In particular to suppliers, some materials and parts require significant specialization and complexity, such as composition requirements or well-defined specifications, which can be maintained through the information stored in RFID tags or using EPCs to map the product to the information stored in a database.

Materials which require constant monitoring of temperature, viscosity, or other physical qualities could also benefit. This also applies to those materials which are heavily time dependent in regards to time-to-disposal, time-to-shipment, and so forth. RFID tags could wirelessly transmit updated information of the state of the material in real time, without human interaction or the costs of installing and/or maintaining an infrastructure based upon a physical connection. While the wireless monitoring devices would require maintenance in case of failure, the overall complexity of the system and of the maintenance would significantly decrease.

Naturally, there are significant costs involved with such implementations of RFID technology in the supplier's world. Compared

Table 1.

	<u>Costs</u>	<u>Benefits</u>
Supplier	Moderate. High for complex systems, such as monitoring devices.	Improved inventory management and control. Improve demand forecasting.
Manufacturer	Moderate, increasing with the complexity of the product or equipment.	Improved inventory management and control. Improve demand forecasting.
Distributor	Moderate. Package-level tagging. High costs when integrating with tracking systems.	More accurate tracking information. En-route location tracking of packages.
Retailer	High. Item-level tagging required for pervasive implementation.	Improved inventory management and control.
Consumer	Variable. Low for luxury items to extremely high for low priced goods.	Reduce stock-outs. Improved shopping experience. Reducing in price.

with manufacturing or distribution, the information necessary to store is uniquely different. The cost of identification tags and tag readers are a common theme among all of the stages in the supply chain—an unavoidable cost. The supplier does have a slight advantage in this regard, in that relatively few tags are necessary in comparison to the other stages in the supply chain. However, if the materials required highly precise specifications or other physical properties, the complexity of the system required to maintain the information could increase the cost exponentially. Monitoring the state of the materials poses even more problems, requiring specialized identification tags attached to sensors. Additionally, simple EPC codes may no longer be sufficient to monitor the possible states of the materials and therefore require a more specialized system for each individual monitoring tag. In this regard, semantic markup languages similar to PML may become incredibly useful in such applications.

Manufacturers

Similar to suppliers, manufactures have much to gain in regards to simplifying inventory management and increasing the robustness and

richness of the inventory system on the whole. However, what is unique to manufactures is the need to maintain large amounts of data on highly specialized or customized parts or products. RFID tags will provide item-level information unique to the particular part or product, which remains with the individual part of product. Manufacturers will find a significant reduction in the effort necessary to manage the inventory of parts or products on hand and find an increase in the accuracy of that inventory.

Manufacturers often require many complex pieces of equipment for specific applications which, in a large bustling factory, may become lost or simply difficult to find. TransAlta found that tagging pieces of equipment across the 600-foot plants made finding and maintaining the equipment easier and more flexible (Malykhina, 2005). Using Wi-Fi and Bluetooth wireless technologies in conjunction with RFID to blanket the entire facility, TransAlta was able to locate equipment regardless of its location in the company's large facilities. Active RFID tags were used to eliminate the need for manual operation and provide real-time information about the location of the equipment and specific metrics from temperature gauges, vibration probes, and a variety of other peripherals (Malykhina, 2005).

If suppliers could have a problem due to complexities in inventory management, manufacturers have an overabundance of them. Implementing RFID throughout the manufacturing process requires that the individual raw materials and parts from other manufacturers be tagged, and the ultimate product to be shipped out the door also be tagged, either individually or as a package. The system becomes exorbitantly more complex when the manufacturing process requires multiple steps where information of the part or product at each stage is required.

Here, at the manufacturing stage, managers will be first posed with a difficult question when considering implementing RFID technology at their site. The question is whether package level or item level will be more economical for the specific application. For larger products or specialized equipment, tagging individual items is an economical choice. However, if the factory produces millions of widgets per month with little or no variation in those widgets, tagging them on an individual level would be a foolish choice. In most cases, manufacturers have little need for individual tagging of parts and products that come off the assembly lines, leaving package-level tagging a more prudent choice.

Distribution

In isolation, distributors are able to see significant benefits. One of the most significant proposed implementations of the technology in distribution channels is the ability to locate in real time each individual shipment, regardless of its location, and to provide information about its shipment location, destination, content, and so forth. This can be accomplished through combining several other mobile technologies. One implementation suggested the use of GPS technology to locate and identify individual

vehicles, then remotely transmit the information obtained by scanning the individual shipment tags, thus providing remote access to the current contents of the vehicle, regardless of its location. Such a system could provide some value to the business and consumers alike. Richer tracking information allows interested parties to know exactly where a highly important package was last scanned, even if the package or shipment is “en route.”

Wireless mobile technologies are used throughout distribution channels for varieties of business benefits, such as geographic positioning systems (GPS), for tracking and monitoring vehicles in distribution channels (Faber, 2001; Schindler, 2003). However, one issue distributors have with current mobile tracking technologies is that en-route information of vehicle contents is almost impossible to monitor, which is where RFID technology has significant promise.

However, several obstacles hinder the practical feasibility of such a system. First, current tracking systems are easily able to provide similar information, but nevertheless lack the remote, real-time tracking such a system would offer. The practical issue comes with the fact that the cost involved with implementing such a system still surpasses any benefit, even considering how the cost of RFID per tag and per reader has become more reasonable.

Retail and Consumers

Similar to suppliers, if the product retail is selling is dependent upon time—that is, perishable with a limited shelf life—those must be sold, moved, or discarded by a particular date. By outfitting each individual item sold within the store, single, outdated items will not find themselves sitting on then shelves for extended periods of time. A simple system could automatically read the tags and inform the employ-

ees which items have or will soon pass the date in question.

There have been several technologies over the years which have been considered as replacement for the aging UPC standard. However, few have provided sufficient benefits over and above UPC labels. Electronic tags were considered as a replacement, but the benefits over the current standard were so minimal that the small cost associated with each tag was too great to justify the switch. On the other hand, RFID tags provide richer information about the product over and above the current UPC standard, which simply identifies the product. Since the RFID tag can be read via a wireless connection, many items can be scanned and identified in seconds, whereas the present UPC standard requires line-of-sight reading through an optical scanner. This application could further be extended into the often-dreamed automatic checkout machines.

Current theft-deterrent technologies are an independent system from the limited UPC standard. The technology is often unreliable, resulting in countless false positives, and requires that the electronic tag passes through a small area before detection. RFID, on the other hand, would be able to provide for both the UPC label functionality discussed previously, as well as a rudimentary theft deterrent system similar to the ones currently used in retail outlets. Through RFID, retailers would be able to know exactly what products are entering and leaving the store, and which have been purchased and which have not. However, cost barriers are significant in this application since item-level tagging is necessary for it to be effective. Until the cost of individual RFID tags drops substantially in comparison to the price or quantity of the product, in the realm of fractions of a cent, it is not likely we will see widespread, item-level application of RFID. However, some retailers have begun tagging larger, higher priced items

for this purpose. Both Wal-Mart and Woolworth's in the United Kingdom have begun tagging items considered high risk, such as CDs, mobile phones, computer accessories, and other electronic goods (Smith & Konsynski, 2003).

However, retail has a particularly difficult time when isolated from the rest of the supply chain in the implementation of RFID technology. Although the obstacles from the distribution side of retail are not insurmountable—requiring the contents of palettes and packages to be tagged at the distribution centers, similar to package-level distribution or manufacturing inventory, before being sent to individual retail stores—on the consumer side however, implementation becomes a much more difficult task. To be effective, individual items must be tagged. While other stages could have survived with a small amount of identification tags and tag readers, retailers cannot avoid the substantial costs associated with the thousands upon thousands of tags to cover individual items in some RFID applications. Inventory management and automatic checkouts require each individual item in the entire store to be appropriately tagged, with no assurance that the tags could be reused. Effectively, the costs of tagging individual items would be directly added to the wholesale cost of each item, and at the current five cent-mark—a substantial portion of many items' wholesale costs—further reducing the already meager margins in retail. Some organizations have implemented theft-prevention systems utilizing RFID technologies, but it is a small fraction of merchandise, particularly high-risk or high-cost items. However, retail systems are particularly simple in comparison to those found in the other stages. Since all items already possess UPC labels, mapping EPC to individual items via RFID tags may not be a daunting task and would provide some technological redundancy in case of system failure.

STREAMLINING THE SUPPLY CHAIN

Implementing technologies in the supply chain ultimately creates value when each organization at each stage of the supply chain vertically integrates, standardizing on the single technology. RFID is no different. Using the same technology from the manufacturing of a product to the sale of the product throughout the entire supply chain substantially reduces costs and provides business value for everyone (Poirier, 1999). However, this does not always occur, for example, when an integral part of the supply chain chooses not to implement or share information, or organizations force their supply or demand chain to implement a particular technology, largely at the cost of those implementing (Kinsella, 2003). In the following section, how RFID can provide value to a more integrated supply chain through the implementation of RFID technologies will be discussed, ultimately culminating on a view of the entire supply chain. While many of the applications, costs, and benefits have been covered in previous sections, how their applications tie together in the supply chain and provide value will be the focus.

Suppliers to Manufacturers

The real power and value of RFID technology in supply chain management sadly comes later in the supply chain, although many benefits can be realized between suppliers and manufacturers. Materials are cultivated, packaged, distributed, received, and processed by the respective manufacturers. Throughout this entire process, RFID is able to track the inventories of materials, the source of the materials (the supplier), and the destination (the manufacturer). Manufacturers then benefit by knowing what inventory they have on hand of pre-production mate-

rials and from what vendors those materials originated. This improves the production process at manufacturing facilities greatly if such information is already accounted for by the source of the materials. However, because of the transforming nature of the manufacturer, the same RFID tags are not valuable to the remainder of the supply chain. Once the materials are transformed into products and finally head downstream to distributors and retailers, they must be retagged. In fact, after every transformation of the product, the nature of the product changes and tags must be reapplied.

Manufacturers to Distributors

Here, at the manufacturing phase, we begin to see the real potential of implementing RFID technologies and the scope of their effect on the integrated supply chain. Package-level tagging at the manufacturing level before distribution occurs both helps maintain inventories at the manufacturing sites and aids distribution and other inventory management and control systems to be equally as streamlined. Incoming materials from suppliers are able to be logged, and inventories updated and maintained throughout the manufacturing process. However, the beauty occurs when item-level tagging is implemented. As the product is produced, the item may be tagged with manufacturing information and other specifications particular to the product. Other organizations down the supply chain will be able to access this information even when the package it was contained in was dismantled and the contents strewn across retail stores and the consumer population. Additionally, the benefits from improved forecasting comes from information downstream, at the retail level, where manufacturers are able to determine which products are sold, at what locations, in specific quantities. For retailers to provide this information, with the RFID

network previously discussed in place, would be far from an insurmountable feat, and the value coming downstream from the manufacturers would be to their benefit. While manufacturers ultimately may bear much of the implementation cost, they will receive equally in benefits, with even more significant benefits for organizations downstream from the manufacturer.

Distributors to Retailers

Leaving the distribution centers are countless cases and pallets of merchandise heading to different retail stores with varying quantities of thousands of different products. Managing what products are leaving or being received and where they are going can become a daunting task, even with some of today's technologies. For distributors, RFID technologies provide some of the more impressive benefits, even in isolation with relatively smaller increases in cost. However, having package-level inventories tagged by suppliers and manufacturers before entering the distribution channels improves the efficiency of the logistics systems for both those parties. As the tagged packages move through the distribution channels, retailers ultimately will benefit as well as the packages move through their receiving centers, actively managing the incoming inventory at individual stores. However, between manufacturing and retail, distributors must retag if the packages themselves are repacked. Luckily, if this is not the case, the RFID network system provides separate semantic information for each of the EPCs associated with the packages. Again, the beauty is at the item level. Since each individual product is now tagged, even repackaging the products does not require new tags to be placed on them. In fact, distributors need not retag any items whatsoever, only change the information associated with the

corresponding EPC, which is unique for all the items entering and leaving the distribution channel, regardless of the location.

Tesco, a large UK retailer, recently implemented an RFID system of significant size, totaling 20,000 identification tags for their stores and distribution centers, with 4,000 tag readers and 16,000 antennas to receive the identification tag signals (Sullivan, 2005). RFID tags have been installed in order to track the merchandise cases and pallets which grace the docks at the distribution centers and receiving doors at retail stores. Unlike the retail giant Wal-Mart, however, Tesco made the investment in RFID themselves independent of their suppliers, in hopes that they perceive similar cost benefits as Tesco.

In some situations, logistics systems may need to make a sudden re-route of product or material in case of a sudden stock-out. When a product is in demand, having no inventory of a product on hand means lost sales for retailers or lost production time for manufacturers. In conjunction with GPS and cellular technologies (Schindler, 2003), distributors now may locate items en-route between destinations and calculate precise inventories of those vehicles. If a reroute is economically reasonable, the vehicle is able to be informed of where the reroute is located and what products the reroute are for. All in all the distribution system becomes more flexible and capable of providing for the retail and customer base.

Retailers to Consumers

One of the significant benefits from implementing RFID technology at the retail level is the reduction of labor costs from managing inventory, which now can be accounted for and monitored with little or no manual operation. Reduction in labor costs provides for two potential outcomes that benefit consumers. First,

the additional labor capacity can be used to improve customer service of the retail establishment for customers. Or, the additional savings in labor costs not rerouted to another activity could be brought directly to the consumer in the form of lower prices, ultimately an increased value for the customers.

Merchandise Security

RFID technologies also move the retail and consumer relationship to the Holy Grail, the market of one. Prada of New York is one of the first retailers to use RFID technologies to revolutionize the shopping experience of its customers. Each item sold in the store is tagged with an RFID tag. Naturally, this provides additional security in the case of stolen merchandise; however, the interesting aspect of the implementation occurs when the customer re-enters the store. Whether carrying or even wearing the previously purchased garment, the tag readers at the entrance to the store scan for an RFID tag; if found, information pertinent to that garment appear on large flat-panel displays around the store. For example, items matching that garment may appear, in reasonable sizes to the item purchases, directing the customer. In addition, richer marking information is obtained through this system, as item purchases are now tied together with how frequently the customer visits the store, what items are purchased in combination or sequences, and so forth. However, consumer privacy concerns have already arisen in regards to this implementation.

From the information gathered here, at the retail and consumer level, suppliers and manufacturers are now better able to forecast demand and control inventories, sending the business value back upstream (Lapide, 2004). The value invested earlier in the supply chain to tag either packages or items leaving manufacturing

facilities returns to them through this improved ability to forecast demand.

CONCLUSION AND FUTURE DIRECTION

RFID technology is a fairly simple wireless technology, composed of a small antenna and microchip, and able to streamline the mobile supply chain. Technologies surrounding the RFID technology, such as EPC and PML, improve the interoperability, transparency, and flexibility of implementing RFID systems with current inventory management and distribution systems. The mobile nature of the technology incorporates additional advantages only found with more complex, higher cost systems. However, important cost considerations must be given, as choosing between the costs and benefits for package level and item level becomes an important decision. While it provides substantial value for organizations downstream, it requires significant investment upstream. As additional implementations appear throughout the supply chain, the cost of the technology will fall and the relative benefits will increase. If standardized on RFID technology, regardless of package- or item-level implementation, the entire supply chain benefits from a standard mechanism to identify objects moving up and down the supply chain, through distribution channels, and off the shelves at retail stores. RFID is poised to revolutionize the supply chain by streamlining operations, providing flexible, transparent communication between organizations.

REFERENCES

Asif, Z., & Mandviwalla, M. (2005). Integrating the supply chain with RFID: A technical and

- business analysis. *Communications of the AIS*, 15, 393-426.
- Angeles, R. (2005). RFID technologies: Supply-chain applications and implementation issues. *Information Systems Management*, 22(1), 51-65.
- Anonymous. (2003). Supply chain technologies—At Woolworth's. *Work Study*, 52, 44-46.
- Atkinson, W. (2004). Tagged: The risks and rewards of RFID technology. *Risk Management*, 51, 12-19.
- Kinsella, B. (2003). The Wal-Mart factor. *Industrial Engineer*, 35, 32-36.
- Lapide, L. (2004). RFID: What's in it for the forecaster. *Journal of Business Forecasting Methods and Systems*, 32(2), 16-19.
- Leary, D. E. O. (2000). Supply chain processes and relationships for electronic commerce. In M. Shaw, R. Blanning, T. Stradder, & A. Whinston (Eds.), *Handbook on electronic commerce* (pp. 431-444). Berlin: Springer-Verlag.
- Malykhina, E. (2005). Active RFID meets Wi-Fi to ease asset tracking. *Information Week*, 1022, p. 38.
- Poirier, C. C. (1999). *Advanced supply chain management*. San Francisco: Berrett-Koehler.
- Rappold, J. (2003). The risks of RFID. *Industrial Engineer*, 35, 37-38.
- Schindler, E. (2003). Business: The 8th layer: Location, location, location. *netWorker*, 7(2), 11-14.
- Smith, H., & Konsynski, B. (2003). Developments in Practice X: Radio frequency identification (RFID)—An Internet for physical objects. *Communications of the AIS*, 12, 301-311.
- Sullivan, L. (2005). UK retailer goes in RFID shopping spree. *Information Week*, 1022, p. 36.
- Yang, B. R. (2000). Supply chain management: Developing visible design rules across organizations. In M. Shaw, R. Blanning, T. Stradder, & A. Whinston (Eds.), *Handbook on electronic commerce* (pp. 445-456). Berlin: Springer-Verlag.

Chapter LXIII

Mobile Batch Tracking—A Breakthrough in Supply Chain Management

Walter Hürster

T-Systems International, Germany

Hartmut Feuchtmüller

T-Systems International, Germany

Thomas Fischer

T-Systems International, Germany

ABSTRACT

Globalization and expanding markets has invariably led to increasingly higher loads of goods traffic. This has resulted, amongst other things, in challenges to supply chain management in terms of cost pressure and demands for short-term availability of the goods. Considering that an increasing number of goods will be “on the road” (on rails, on ship, in the air) for an appreciable percentage of the life-cycle, there is an urgent need to bridge the information gap between the automated systems at the factory sites and the storage control systems at the destination sites. This chapter reports on a system solution that has been developed by T-Systems’ Solution and Service Center Ulm / Germany, within the Service Offering Portfolio “Embedded Functions”. The system solution has been gained as a synergy effect of connecting mobile communication solutions with Auto ID Services. It is presented here and discussed in the context of online surveillance during transportation, providing both downstream batch tracking, as well as upstream traceability.

INTRODUCTION AND BACKGROUND

Traditional problems of managing resources and the flow of material appear to have been

solved by enterprise resource planning (ERP) systems as well supply chain management (SCM). This is true of the stationary case of an isolated factory and of the goods that form part of its inventory. However, with the increasing

movement of goods, a new dimension of problems has arisen that makes it inevitable to consider transport status itself—particularly to improve the supply chain planning and the execution process. This chapter is an attempt to cope with the new challenges that result from a higher degree of mobility, a higher percentage of the mobility phase with respect to the total lifecycle, and a higher flexibility with respect to transport media and changes of the transport mode within one single transaction, such as conveying a pallet from A to B (where A and B may be located anywhere on the surface of the earth, thus indicating that also increasing distances have to be bridged).

Goods, spare parts, and assembly components are no longer kept in storage for long periods of time, but are fed in when needed. This is the effect of the popular just-in-time (JIT) approach to inventory management. Thus, managing the supply chain effectively means managing more and more of the transportation chain.

Successful attempts have been made to manage the internal transport at a factory site by means of new technologies, such as radio frequency identification (RFID) tagging or other auto ID technologies (ten Hompel & Lange, 2005). Within this context, a new class of middleware is emerging, acting as a platform for managing the data and routing them between tag readers and enterprise systems (Leaver, 2004). However, a huge gap of information exists for the increasing time of external transportation—either between two factory sites for a semi-product or between factory site and end user location for a final product.

THE CORE CHALLENGE IN SUPPLY CHAIN MANAGEMENT

In order to obtain an exact overview at any time, it is essential to track the flow of goods on batch level at least, if not on item level (for

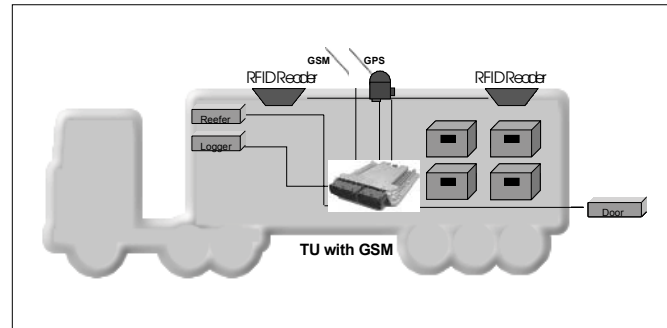
larger items). This requires acquiring knowledge about the geographical position whenever needed plus detailed information about the goods—that is, batch identification and batch description, including information about origin and destination, plus all intermediate agents involved in the process. Regulation (EC) No. 178/2002 of the European Parliament and of the Council of January 28, 2002, as an example, is laying down the general principles and requirements of food law and at the same time the procedures in matters of food safety. This includes strong implications with respect to downstream trackability (from origin to destination), as well as upstream traceability (from end user back to the production site). In the case of non-preservative food, it is of essential importance to monitor and to record the environmental data of the transport—for example, to ensure that the refrigerating chain has not been interrupted (or only for a very short period of time and within a certain temperature range). The big challenge therefore consists of getting all the required information while the goods are on their way on a transport medium in motion.

THE SOLUTION TO THE CHALLENGE

The requirements mentioned above directly lead to the way of finding an appropriate solution by a decomposition of the system into its two basic components:

- a. Subsystem to determine the geographical position of the transport medium (container, lorry, trailer, wagon, ship, aircraft, etc.).
- b. Subsystem to gain information about the goods transported by that medium—that is, batch identification and batch description (plus additional environmental parameters).

Figure 1. Trailer equipment



The first subsystem (a) preferably consists of a GPS antenna and a GPS receiver to obtain the geographical position. For the second subsystem (b), an advanced approach would be to use RFID technology—that is, RFID tags affixed to the packaging units and RFID readers installed on the transport medium to read the tags. An example is shown in Figure 1 (for the case of a trailer/lorry configuration).

The trailer contains the GPS equipment plus RFID reader(s) to identify and to read the tags which are fixed at the package units. A great advantage of RFID vs. other auto ID technologies is due to the fact that no direct geometrical line-of-sight between tag and reader is required—that is, the packages may be oriented in any arbitrary way and do not have to be aligned or rotated in a specific manner. Addi-

tional environmental parameters, such as temperature, acceleration (shock), door status, and so on (including intrusion alarm) are polled from adequate sensors by a so-called reefer and are stored locally on a data logger. All data are collected online by a telematic unit (TU) and are transmitted instantaneously or at given time intervals to a Transport Tracking Center (TTC), preferably by means of GSM or by using satellite communication (depending on the coverage and the location of the transport medium on its route). This Transport Tracking Center, thus defining the third subsystem (c), collects the batch data from all connected transport units and makes them available to all subscribers and stakeholders being entitled to use them. The TTC itself will consist of a computer cluster with distributed tasks for I/O handling, central

Figure 2. Network of information

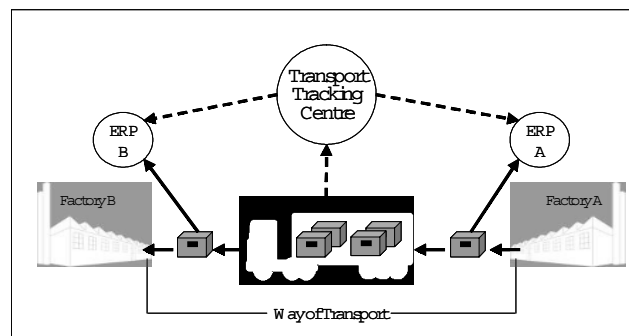
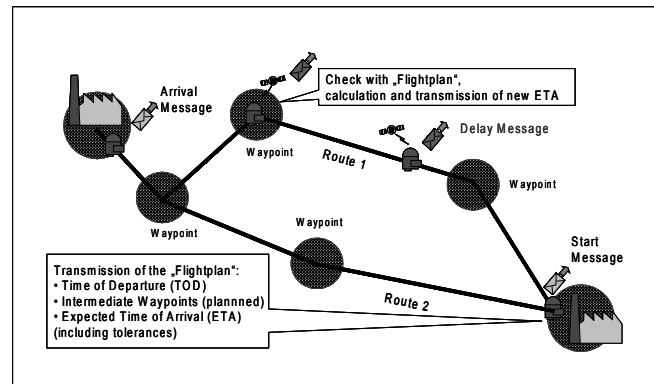


Figure 3. Intermediate waypoints



data storage, archiving, and data retrieval. Figure 2 shows an example for this network of information, again for the lorry/trailer configuration mentioned above:

Whenever a batch is leaving Factory A, a stationary tag reader identifies the batch, and the batch data are transmitted by an appropriate middleware to the ERP system of Factory A. The same procedure will take place when the batch arrives at Factory B (or at the site of final destination). The full information is available at any time by connecting the stationary ERP system to the Transport Tracking Center by means of a proprietary telecommunication link or via Internet. By those means, it is possible to obtain a more reliable estimate of the time of arrival for a specific good—thus allowing for a rearrangement of the production line at the destination site within due time (if necessary).

The eminent advantages of such a system become increasingly obvious if the transport is not a single point-to-point connection, but if a number of intermediate waypoints have to be covered, including unloading of some batches and loading of new batches, as illustrated in Figure 3, and whenever the transport medium is changed (e.g., from lorry to train) and a new batch configuration has to be assembled.

The system described here allows the user to gain an exact overview at any time and to track the flow of goods from the origin to the destination online and in real time. On the other hand, by means of archiving and retrieval, it allows for backward (upstream) traceability at batch level (or item level). In other words, by closing the mobility gap, this system is covering the full supply chain without any interruptions.

T-Systems has implemented such a system (called “eCargo”) for RAILION, Europe’s largest international logistics enterprise for railway-based transports, as described by Epple and Feuchtmüller (2005). More than 100,000 individual RAILION transports per day are crossing all over the continent, carrying a huge variety of goods. About 13,000 wagons are equipped with GPS and GSM devices at least, plus environmental sensors, reefers, and data loggers.

For a pilot installation and for operational use, a similar system (called “iTM”®—intelligent Tracking Management) has been developed by T-Systems for Schmitz CargoBull, one of Europe’s leading trailer manufacturers. Reliability, safety, and security are top priority requirements for both systems.

Table 1 provides a summary of the benefits and advantages of the described system solution.

Table 1. Benefits and advantages of mobile batch tracking systems

Commercial Aspects	Security Aspects
• Fulfilling the requirements with respect to batch tracking	• Secure authentication by full time coverage and uninterrupted data history
• Timely implementation of EU Regulation 178/2002 with respect to food batches	• Basis for certification according to IFS (International Food Standard)
• Online trackability (“downstream”)	• High security by “closed-door” principle (reliable content management)
• Data retrieval and traceability (“upstream”)	• Documentation of the grower, producer, or manufacturer
• High degree of automation by using RFID technology	• Documentation of the receiver/user and of the intermediate agents
• Minimization of damage in case of recall actions	• Documentation of the wares, the raw materials used, and of all relevant time stamps
• Interface to customer ERP systems	• Documentation of the environmental conditions during transportation
• High degree of intermodal flexibility	• Documentation about storage and status of semi-products and intermediate processing stages

LIMITATIONS AND CHALLENGES TO THE SOLUTION

Basically, there are no other limitations to a worldwide use of the system than those imposed by physics. Perturbations of the radio frequency (RF) may have to be faced if RFID is used in a ferro-metallic environment, or Faraday screening may prevent readers from identifying tags if they are “hidden” by a metallic foil. A further electromagnetic threat is encountered if the system is operated in environments with spark discharges or in cases of other events causing high-voltage electromagnetic pulses (EMPs). In those cases, the transponder chip may be completely damaged.

Besides these physical challenges and threats, a global use is rather endangered by incompatible or even competing systems with respect to the performance of readers and transponders, and with respect to the RF used. While in the U.S. the UHF range between 868 and 915 MHz is favored, many developments in

Germany prefer a frequency of 13.56 MHz (ten Hompel & Lange, 2005).

A different limitation to an increasing use of RFID-based systems may be given by a more commercial point of view. As for all new technologies in the beginning, the unit prices are relatively high (approximately 0.60 USD per transponder tag, depending on the storage capacity, ranging from a few bytes to several Kbytes). Drastic price reductions can be expected for the time to come when large numbers of tags will be produced.

Finally, like in many other cases, it has to be considered that the system is subject to some security risks and to the possibility of criminal attacks. Removing or destroying tags by brute force is the simplest way, followed by more sophisticated acts such as unauthorized reading of the tags, cloning of tags by means of electronic devices, and emulation of tags with any desired content (Oertel et al., 2005).

Here again, further development of the technology and international security standards will help to reduce the inherent risks.

CONCLUSION AND FUTURE DIRECTION

This chapter has shown that a system solution for mobile batch tracking is feasible that allows for online batch tracking during downstream transportation, as well as for upstream traceability. The system presented here bridges the information gap between the automated systems at the factory sites and the storage control systems at the destination sites. By using finest technology according to the state of the art, this mobility system can be considered to represent a breakthrough in supply chain management—especially when taking into account that an increasing number of goods will be “on the road” (on rails, on ship, in the air) for an

appreciable percentage of the lifecycle, thus resulting in an urgent need to cover this mobility phase.

Nevertheless, a number of problems and difficulties still persist. Due to the international nature of the system, it is quite obvious that full functionality across borders will require international agreements, legal regulations, and standards. Technological standards will have to deal with reserved frequency ranges for the RFID equipment, the transmission speed, coding, protocols, and anti-collision procedures. Data standards will have to take care of a scheme for unique numbering (e.g., according to ISO/IEC Standard 15963), and application standards will have to consider new coding standards, such as the Electronic Product Code (EPC) replacing the UPC Barcode Standard (ten Hompel & Lange, 2005). Unique identification will require a well-elaborated coding standard based on a worldwide agreement, especially when thinking in terms of progressing from unit- and pallet-tagging down to item-tagging.

In parallel, sophisticated security measures will have to be developed in order to overcome the criminal risks inherent to each new technology.

Provided that these prerequisites are given, there is no doubt that mobile batch tracking systems based on RFID technology—like the one presented here—will result in a tremendous improvement of supply chain management.

REFERENCES

- Epple, M., & Feuchtmüller, H. (2005, January). Weichen für die transportsicherheit (Points for the safety of transportation). *Europäische Sicherheit (European Safety)*, (1), 56-57.
- Leaver, S. (2004, August 13). *Evaluating RFID middleware* (Company Research Report), Forrester Tech Choices, Forrester Research, Inc., Cambridge, MA, USA.
- Oertel, B., Wölk, M., Hilty, L., Kelter, H., Ullmann, M., & Wittmann, S. (2005). Der gläserne kunde (The glassy customer). *DoQ, H&T Verlag*, (1), 53-55.
- ten Hompel, M., & Lange, V. (2005). Barcode geknackt (Barcode cracked). *DoQ, H&T Verlag*, (1), 48-50. München, Germany.
- T-Systems. (2005). Retrieved April 12, 2005, from www.t-systems.com

About the Authors

Bhuvan Unhelkar has 24 years of strategic as well as hands-on professional experience in information and communication technology. Founder of MethodScience.com, he also has notable consulting and training expertise in software engineering (modeling, processes, and quality), enterprise globalization, Web services, and mobile technologies. He earned his doctorate in the area of “object orientation” from the University of Technology, Sydney. In his academic role at the University of Western Sydney, he leads the Mobile Internet Research and Applications Group (MIRAG), has authored/edited eight books, and has extensively presented and published research papers and case studies. He is a sought-after orator, a fellow of the Australian Computer Society, a Rotarian, and a previous TiE mentor.

* * *

Christopher Abood has worked in ICT for over 20 years with a Master of Commerce degree, majoring in Information Systems, from the University of New South Wales. He is an active member of the Australian Computer Society; during 2004, he led the development of the Mobile Camera Phone policy for the Australian Computer Society, and has done a number of newspaper and radio interviews on the inappropriate use of mobile camera phones.

Harpreet Alag is currently working as a senior business process consultant with Agilisys Limited, UK. He has a number of years of experience in process change and business analysis for

developing and implementing business systems in multiple domains and business areas. In the last few years, Mr. Alag has been involved in major change programs, implementing CRM and ERP applications along with business integration. The focus of his work has been business process redesigning and implementation of application packages, and he has also been involved in change and transition management.

Chantal Ammi is a professor at the National Institute of Telecommunications, Paris, France. Specializing in marketing and strategy in high-tech sectors, she has published several books in different fields, has worked on different international contracts of research, and has developed different products and services.

Dinesh Arunatileka is researching for his PhD at the University of Western Sydney, Australia, in the area of mobile technologies and their application to business processes. He has also been a teaching fellow at the same university for the past two-and-a-half years. He earned his BSc in Computer Science from the University of Colombo and his MBA from the University of Sri Jayewardenepura, Sri Lanka. He has over nine years of experience in business development in the computing and telecommunications industry. He has published and presented conference papers in the area of methodologies to introduce mobile technology into business practices.

Achraf Ayadi is PhD student in Management Science at the National Institute of Telecommunications, Paris, France. He has published several articles on banking technologies management, electronic and mobile financial services, entrepreneurship, and Science Parks creation. He is currently the chairman of the PhD graduates and PhD students association of INT (Doc'INT) and associate general-secretary of the Tunisian association of high-schools graduates (ATUGE).

R. M. Banakar received a BE degree in electronics and communication engineering from Karnatak University, India, in 1984 and an MTech in Digital Communication from Regional Engineering College, Surathkal, Karnataka. She has a couple of years experience with the Indian Space Research Organization (ISRO). She completed her PhD in the area of low-power, application-specific design methodology from IIT Delhi in 2004. Presently she is working as an assistant professor at the engineering college, Hubli Karnataka. She is the member of the ISTE, IETE, MIE, and IEEE scientific and professional societies. Her current areas of research include SOC, VLSI architecture, and WCDMA. She can be reached via e-mail at banakar@bvb.edu.

Franck Barbier is a professor of software engineering at the University of Pau, France. Previously, he was the director of the Computer Science Research Institute of the University of Pau from October 2000 to October 2004. His main research focuses are software component modeling, evaluation, distribution, and mobility in the context of UML. He was the scientific consultant of Reich Technologies, a French company among the 17 companies that built UML 1.1 at the Object Management Group in 1997.

Francisco Barcelo earned a degree in telecommunications engineering and a PhD from the Technical University of Catalonia (UPC) in 1986 and 1997 respectively. In 1987 he joined the School

About the Authors

of Telecommunications Engineering of Barcelona at UPC, where he has been teaching design and planning of communication networks. After graduation, he did research in the areas of digital network synchronization and switching. Since 1997 he has been an associate professor at UPC. He also serves as a consultant to the telecommunications industry and operators in Spain, and is currently involved with several research projects supported by the Spanish Government (Plan Nacional de I+D) and the European Commission (IST 5th Framework Program). His current research interests lie in the study of the evaluation and planning of the capacity of wireless networks and in the area of location technologies for cellular and WLAN networks.

Joseph Barjis is currently working as an assistant professor in the department of IT at Georgia Southern University, USA. He is also chairman of the research committee in this department and a member of the Faculty Research Committee at the university level. He earned BSc, MSc, and PhD degrees in computer science. He is actively conducting research on modeling and simulation of business processes, information systems design, and related topics. He has published more than 70 conference and journal papers, four book chapters, and one edited proceedings.

Raghunadh K. Bhattar received his BSc (Engg) from Regional Engineering College, Kurukshetra, India, and his MSc (Engg) from Indian Institute Science (IISc), Bangalore, India, in 1985 and 2000, respectively. Presently he is pursuing a PhD degree in Electrical Engineering IISc. He joined the Space Applications Centre, ISRO in 1990, where he is actively involved in the design and development of Turbo codec on FPGA for ground-based SATCOM systems and contributed to MPEG2 video codec, JPEG, and JPEG2000 systems. His current research interests are digital signal processing, image and video coding, multimedia, wavelets, channel coding, and wireless technologies.

Mohamed Boulmalf received his BSc in Communications from the National Institute of Telecommunications in 1987, Rabat, Morocco. He worked for five years as a network engineer at ONCF Company, Rabat, Morocco. In August 1992, Dr. Boulmalf moved to Canada to pursue his graduate studies at the National Institute for Scientific Research, Montreal, Canada. He received MSc and PhD degrees, both in wireless communications and networking, in 1994 and 2001, respectively. From September 1994 to December 1998, he was with INRS-Telecom as a radio communications research engineer. In January 1999, he joined ETS, Quebec University where he worked as a lecturer. In September of the same year, Dr. Boulmalf moved to Microcell Telecommunications, GSM Operator in Canada, where he worked as a senior network engineer. From 2000 to 2002, he worked as a principal engineer in the Multi-Vendor Integration Department at Ericsson, Montreal, Canada. In February 2002, he joined the College of Information Technology at the United Arab Emirates University, Abu Dhabi–Al-Ain, where he is now an assistant professor.

Matthias Brantner studied Information Systems at the University of Mannheim, Germany, from 1999 until 2004. He is currently employed and working on his doctorate at the Chair of Practical Information Technology III in Mannheim. The topic of his doctorate is query evaluation of XML query and transformation languages.

Torsten Brodt joined the Mcminstitute at the University of St. Gallen as a PhD researcher in 2003 specializing in mobile communication and media-related research. He is leading international research projects with MNOs, media, and high-tech companies. From 2001 to 2003 he worked as a consultant at A.T. Kearney management consultants and gained substantial experience in the European mobile communication industry. Mr. Torsten holds master's degrees in management science from the University of Mannheim, Germany, and Trinity College, Dublin, Ireland. His research interests include the market-orientation of innovation networks, and user-centric product and service development.

Narottam Chand received a BTech degree from the National Institute of Technology, Hamirpur, India, and an MTech degree from the Indian Institute of Technology, Delhi, in Computer Science & Engineering. He is currently a research scholar in the Electronics & Computer Engineering Department at the Indian Institute of Technology, Roorkee. His research interests include mobile computing and ad hoc networks. His PhD research work continues in the subject of data dissemination and caching techniques in mobile computing environments.

Chris Chatwin, under the auspices of the Industrial Informatics and Manufacturing Systems Research Centre (IIMS), has successfully completed a number of EPSRC, European, DTI, DFT, and industrial research projects in e-commerce, electro-optics, optical computing, control and systems integration, digital electronics, and image processing. Professor Chatwin has published two research level books: one on new numerical methods for simulation of laser materials processing, and the other on hybrid optical/digital computing. He has also published more than 200 international papers.

Xiao Chen holds a Bachelor's of Computing (E-Commerce) degree from the University of Western Sydney from their joint course run at Nanjing University of Chinese Medicine (NUCM), China. He is currently studying for his master's degree as an international student at Monash University, Australia. He has had special training and work experience in enterprises such as Siemens AD and Gori in China, enabling him to form a good understanding of mobile technologies in China.

Andrew P. Ciganeck is approaching the final stages of his PhD program at the University of Wisconsin-Milwaukee. His research focus is on the adoption, diffusion, and implementation of Web services and other Net-enabling initiatives, mobile computing devices, knowledge management systems, and the role that the social context of an organization has in information systems research. He has published in a number of refereed conference proceedings and has manuscripts under various stages of review with scholarly IS journals.

Fred Claret-Tournier received his engineering degree from the School of Electrical Engineers, Grenoble, France, in 1996. In 1997 he received a master of science degree from the University of Sussex, Brighton, UK. In 1998 he joined the Department of Engineering, University of Sussex, as a research fellow. His research interests are in digital image processing, digital signal processing, embedded systems, manufacturing systems, and interface programming.

About the Authors

Joanne Marie Curry joined the University of Western Sydney (UWS) in 1996 specializing in the teaching of information systems. She has been the unit coordinator of the final-year Computing Project units for the last five years. In 2004, she was recognized for her contributions to practical project work, being the winner of the UWS 2004 Vice Chancellors Award for Excellence in the area of Regional and Community Engagement. Ms. Curry established the “XML and the Semantic Web^{3/4}Implications and Applications” mini-track at HICSS in 2001, and is also on the Organizing Committees for the 2005 Web Engineering Conference and the 2006 International Conference on Information Management and Business. She is currently researching for her PhD studies.

Ritanjan Das is a lecturer in information systems in the University of Portsmouth Business School. His research interests include e-business and e-commerce/m-commerce, information systems failure, and information systems security management.

Sipra Das(bit) received her BE in electronics and telecommunication and ME degree in electronics and telecommunication (specialization in computer science) in 1984 and 1986, respectively, and her PhD from the Department of Computer Science & Engineering in 1997. Since 1988 she has been with Department of Computer Science & Technology, Bengal Engineering and Science University (formerly Bengal Engineering College), where currently she is an assistant professor. She is also the recipient of the Career Award for Young Teachers from the All India Council of Technical Education.

K. S. Dasgupta received bachelor's and postgraduate degrees in computer science, with Honors, from Jadavpur University, India, in 1972 and 1973, respectively. He received his PhD from the Indian Institute of Technology, Mumbai, India, in 1990. He joined the Space Applications Centre, ISRO, India, in 1974, and since then has contributed significantly in the field of image processing and satellite communications. As group director of the Advanced Digital Communication Technology Group (ADCTG), he was instrumental in the design and development of a PC-based multimedia system for satellite-based distance education. His current areas of research interest address digital signal processing, digital image processing, computer architecture, and digital communications. He is a senior member of IEEE and a fellow of IETE India.

Samir El-Masri earned an engineering degree in electronics from the Lebanese University in 1993. Dr. El-Masri holds a master's degree (1994) and a PhD in Speech Production from the Institut National Polytechnique de Grenoble (INPG), France. He worked on building a real talking robot at Hokkaido University and NTT in Japan from 1998 to 2001. From 2001-2003, He was a lecturer at Central Queensland University and the University of Southern Queensland. Currently Dr. El-Masri is a senior lecturer at the University of Western Sydney, Australia. His current research interests are Mobile Web Services.

Francesco Falcone started as a project manager in 1986, developing ICT solutions for Banks. After a short move to the Fiat Group in 1989, in 1991 he was with the U.S. system integrator, EDS, and led the development of Internet applications. In 1999 he was named CTO of Digital Business, leader in the development of VAS applications based on new emerging mobile Internet technologies,

partnering with Nokia and Openwave. Leading the Research & Development Department, he managed important projects with Vodafone and Telecom in the mobile VAS, messaging, and IP multimedia environment. He currently is a technology manager at Nok Services Ltd., London.

Hartmut Feuchtmüller (Dipl.-Ing. technical computer science) is a senior consultant for manufacturing services and intelligent tracking management systems at T-Systems International. He is experienced in software project management and consulting for telematics projects and has expertise in business processes for commercial and railway telematic systems, tracking and tracing systems, auto-identification, supply chain event management, and RFID technologies.

Thomas Fischer (Dipl.-Ing. aerospace) is currently head of manufacturing services at T-Systems International. He is experienced in software project management and executive management, and he has gained expertise in navigation systems and avionic systems, tracking and tracing systems, rail control systems, auto-identification, and supply chain event management.

Jason Gan specializes in the software design and development of Web front-ends to integrated business systems. He has 10+ years of experience as a layout artist and currently works as a Web designer and programmer in the industry, developing Web front-ends against Microsoft Great Plains ERP, Microsoft CRM, and Microsoft SQL Server. Mr. Gan graduated from Macquarie University in 1994 with a major in English Literature (Literary Craftsmanship). This was followed by a master's degree in computing science from UTS, Sydney, and he is currently working to produce "Great Plains C#" developer documentation.

Marco Garito was born in Milan, Italy, in 1967. He graduated with a law degree from the University of Milan and started his professional development in a professional and telecommunication services company before entering the Internet world with an incubator, where he stayed until late 2001. He then moved to Sydney to study for a master's degree at the University of Technology. After returning to Europe, he worked with IBM in Scotland (2004) and is currently with Cisco in Northern Ireland dealing with marketing and CRM.

Abbass Ghanbary holds a bachelor's degree in applied science, Honors, and is currently undertaking PhD research at the University of Western Sydney (UWS), Australia. His specific research focus includes the issues and challenges in incorporating Web services in businesses with the aid of mobile technologies and the subsequent issues of business process reengineering. Mr. Ghanbary earned a scholarship from the University of Western Sydney to undertake his research on the effects of and how to improve Web services with the aid of mobility. He also teaches and tutors at UWS, and is a full member of the Australian Computer Society.

Nina Godbole has more than 12 years of experience in the IT industry in system analysis, business development, support services, quality management, and training. She holds a key managerial position at IBM Global Services (India) Ltd., wherein she has successfully driven organization initiatives such as the P-CMM, CMM-I, BS 7799, and participated in the security/IT audits, as well as researching in them. She is author of a software quality assurance professional book (published

About the Authors

in 2004 by NAROSA in India^{3/4}www.narosa.com^{3/4}and distributed in the U.S. and UK by Alpha Science). Ms. Godbole has also presented a number of papers on this topic in national and international conferences. She holds an MS degree from India's reputed Indian Institute of Technology (IIT-Bombay) and also earned an MEngg (computer science) from Newport University, USA. Her professional certifications include CQA, CISA, PMP, and CSTE.

John Goh is a PhD student at the School of Business Systems, Faculty of Information Technology, Monash University, Australia. He is conducting research in the field of mobile data mining, which focuses on extracting interesting patterns and knowledge from raw data collected from mobile users, including mobile phones and personal digital assistants. Since his enrolment into a research degree in late 2003, he has published a number of papers in the area of mobile data mining.

C. Gomathy acquired a BE (Honors) in electronics and communication engineering from the Government College of Engineering, Tirunelveli, in the year 1986, and an MS in electronics and control engineering from the Birla Institute of Science and Technology, Pilani, in 1992. She also obtained an MS (by research) from Anna University in 2001. She is currently pursuing her PhD in the Department of Electronics and Communication Engineering, College of Engineering, Anna University, Chennai, India. She has published more than 18 research papers in national and international conferences and journals. Her areas of interest include mobile ad hoc networks, high-speed networks, and digital communication.

Sheng-Wei Guan received his MSc and PhD from the University of North Carolina at Chapel Hill, USA. He is a chair professor with the School of Engineering and Design at Brunel University. Dr. Guan has worked in a prestigious R&D organization for several years, serving as a design engineer, project leader, and manager. After leaving the industry, he joined Yuan-Ze University in Taiwan for three-and-a-half years. He served as deputy director for the Computing Center and the chairman for the Department of Information & Communication Technology.

A. Hameurlain is a professor with Paul Sabatier University; his main research interests include: parallel databases, mobile databases, mobility, distributed systems, and query response time optimization.

Robert Harmon is professor of marketing and technology management at Portland State University, USA. His research includes value-based marketing, new product development, e-business, mobile commerce, strategic pricing, and all phases of the strategic market planning process. He has published in such journals as the *Journal of Marketing Research*, *Journal of Marketing*, *Journal of Advertising*, *Journal of Advertising Research*, and *Decision Sciences*, among others. He is currently a principal investigator for a \$250,000 National Science Foundation grant focused on value-based software engineering for wireless thin-client applications.

Paul Hawking is a senior lecturer in the School of Information Systems at Victoria University, Melbourne, Australia. He is the SAP academic program director for the Faculty of Business and Law, and is responsible for the facilitation of ERP education and research across the university. Accordingly he is coordinator of the university's ERP Research Group

(www.businessandlaw.vu.edu.au/sap/research.html). He is past chairman of the SAP Australian User Group, and is now responsible for education and research for this group. This has provided him with strong links with SAP customers. Professor Hawking is a leading researcher in the area of ERP systems and has produced many research publications, including a number commissioned by the SAP user community.

Sven Helmer studied computer science at the University of Karlsruhe in Germany from 1989 until 1995. Following that, he acquired a PhD in 2000 doing research in the area of database performance at the University of Mannheim, Germany. He stayed there until 2005, working in the area of native XML database systems as an assistant professor. In October 2005 he joined Birkbeck College, London, as a lecturer. He has published more than 30 papers in various journals, conference proceedings, and books. Furthermore, he served as a reviewer for different journals and as a member of several program committees.

Wen-Chen Hu earned BE, ME, MS, and PhD degrees, all in computer science, from Tamkang University, Taiwan, National Central University, Taiwan, the University of Iowa, USA, and the University of Florida, Gainesville, USA, in 1984, 1986, 1993, and 1998, respectively. He is currently an assistant professor in the Department of Computer Science, University of North Dakota, USA. Dr. Hu has published more than 30 articles in refereed journals, conferences, books, and encyclopedias, and one book, titled *Advances in Security and Payment Methods for Mobile Commerce*. His current research interests are in electronic and mobile commerce, Web technologies, and databases. He is a member of the IEEE Computer Society, ACM, and the Information Resources Management Association.

Walter Hürster holds a PhD in physics and is a principal consultant at T-Systems International; he has been involved in research work in nuclear and particle physics (Germany and Switzerland). He has extensive experience in software project management (large projects) and executive management, expertise in radar systems and avionic systems, radiation protection, coastal protection, remote monitoring of nuclear power plants, early warning and risk management systems, tracking systems, and air traffic control. He has presented numerous papers at international conferences and has contributed to several books on environmental informatics. Dr. Hürster has been called on several times by the EU Commission as an expert for risk management and early warning systems. He is also a member of several organizations and national working groups.

Mohammad Mahfuzul Islam received his BEngg (Honors) and MEngg degrees in computer science and engineering (CSE) from the Bangladesh University of Engineering and Technology (BUET) in 1997 and 2000, respectively. He is currently an assistant professor of CSE at BUET, and is close to completing his PhD at the Gippsland School of Computing and IT, Monash University, Australia. His research interests are mobility support resource management for cellular multimedia networks, genetic algorithms, neural networks, and fuzzy systems. He has published 18 journal and peer-reviewed research publications.

Bradley Johnstone spent many years working as an account executive in the advertising industry in Australia and the United Kingdom. After completing a computer science degree, he

About the Authors

changed careers and is now working in the banking and financial industry. His research interests include application of mobility to banking business processes.

Matthew R. Jones lectures in information management in the Department of Engineering and the Judge Institute of Management at the University of Cambridge, UK. His research interests concern the relationship between information systems and social and organizational change. He has published widely in this area, including several studies in the health care domain.

R. C. Joshi received the BE degree from Allahabad University, India, and ME and PhD degree in electronics & computer engineering from the Indian Institute of Technology, Roorkee. Currently he is a professor in the Department of Electronics & Computer Engineering, Indian Institute of Technology, Roorkee. He has served as chair for various international conferences, including Expert System and Robotics at Pittsburgh (1994), Parallel Processing at St. Charles (1994), and ADCOM at IIT Roorkee (1999). His current research interests include database systems, data mining and knowledge discovery, mobile and distributed computing, sensor networks, and security.

Carl-Christian Kanne studied computer science at the Rheinisch-Westfälische Technische Hochschule (RWTH) Aachen in Germany from 1992 until 1998. In 2003, he acquired a PhD from the University of Mannheim, based on his research on native XML database systems. He is a founder of “data ex machina GmbH,” a German company whose products include the Natix XML DBMS. He has published more than a dozen papers in various journals and conference proceedings.

Heikki Karjaluo is a research professor at the Faculty of Economics and Business Administration, University of Oulu, Finland. His research interests concern electronic business in general, and mobile business and mobile commerce in particular. He has published extensively on electronic business in marketing and information system journals, and has collaborated with several researchers in Finland and abroad and with Finnish high-tech companies in common research projects.

Anand Kuppuswami is a research associate pursuing his PhD in the School of Computing and IT at the University of Western Sydney. He received his MS (Honors) from the University of Western Sydney and his BE from the University of Madras. His primary research interest lies in the application of neural networks and agent technology in the field of pattern recognition and intelligent migration.

Yi-chen Lan is a senior lecturer in the School of Computing and IT, the University of Western Sydney. Dr. Lan holds a Bachelor’s of Commerce—Computing and Information Systems (Honors) degree and a PhD from the University of Western Sydney. He teaches information systems and management courses at both the undergraduate and graduate levels. Prior to his current academic work, Dr. Lan served industry for five years, wherein he held senior management responsibilities in the areas of information systems and quality assurance programs in a multinational organization. His main areas of research include global transition process, global information systems management issues, globalization framework development, integrated supply chain development, and health-related information systems development and management.

Sandra Synthia Lazarus is currently completing her PhD in Medical Informatics at the University of Sydney in which she hopes to evaluate the impact of information technology on clinical processes. As part of her research, she has evaluated wireless technology in a hospital environment. Her previous contributions include research in biochemistry, inhabiting fungal actives, and agile methodologies in computer science. She has completed a bachelor of science degree, a master's in computer science, and hopes to continue her education, while working with various e-health companies to develop e-health solutions.

Thomas Leary is a doctoral student in the Information Systems and Operations Management (ISOM) Department at the Bryan School of Business and Economics at the University of North Carolina at Greensboro, USA. His research interests include mobile technologies, Semantic Web, mobile e-business and strategy, service descriptions and composition, and intelligent multi-agent architectures.

Chean Lee holds a master's in information technology from the University of Western Sydney and a Bachelor's of Commerce from Griffith University. He has been working throughout e-business, Internet, Web development, and ERP application industries. His current research interests include application of mobile technologies to Customer Relationship Management applications.

Maria Ruey-Yuan Lee is the department head and an associate professor in the Information Management Department, Shih Chien University, Taipei, Taiwan. Her current research interests include ontology, wireless Internet applications, and mobile commerce. She has published in a number of international conferences and journals, and has been referee, panelist, program committee member, and organizer at many international conferences. She worked at CSIRO Mathematical and Information Sciences, Sydney, Australia, for more than 12 years before she joined academics in Taiwan. She led a group conducting research and development in applying artificial intelligence technologies to electronic business applications while at CSIRO.

Pouwan Lei is a lecturer at the University of Bradford, UK. Previously she was a lecturer at the University of Macao. Her research interests include electronic commerce, mobile commerce, semantic Web, ubiquitous computing, and intelligent agents. She holds a Doctor's of Philosophy degree from the University of Sussex, UK.

Chye-Huang Leow is a lecturer with the School of Business at the Singapore Polytechnic. She has lectured on subjects including retail environment and technology, organizational management, management and organizational behavior, and services marketing. Ms. Leow has assisted with several key projects for retailers in Singapore in the area of successful business strategies. Her key area of research is business-related and technology applications to education institutions and businesses. Prior to joining the Polytechnic, Ms. Leow was in marketing and retailing practice with local and international companies for many years, providing technical and professional advice to retailers and commercial companies. She is also a certified Casetrust auditor in Singapore.

About the Authors

Matti Leppäniemi is a researcher with the Faculty of Economics and Business Administration, University of Oulu, Finland. He is interested in the areas of mobile marketing and integrated marketing communications.

Feng Li is chair of E-Business Development at the University of Newcastle upon Tyne, UK. His research centrally focuses on the interactions between information systems and emerging strategies, business models, and organizational designs. He has worked closely with organizations in banking, telecommunications, manufacturing, retailing, and electronics, as well as the public sectors. He is the chair of the E-Business & E-Government Special Interest Group (SIG) in BAM and a member of the BAM Council. His work on Internet banking and on telecom value networks and pricing models has been extensively reported by the media. He can be via e-mail at Feng.li@ncl.ac.uk.

Wei Liu holds a Bachelor's of Computing (E-Commerce) degree from the University of Western Sydney from their joint course run at Nanjing University of Chinese Medicine (NUCM), China. Having studied information-centric technologies, he started researching Internet technology, mobile technology, and supply chain management during his final year at NUCM. His research interests include the application of mobile technologies to business processes.

M. Mammeri is a professor at the Paul Sabatier University. His main research interests include: real-time systems, distributed applications, wireless networks, and mobile environments.

Ioakim (Makis) Marmaridis holds a bachelor's in business computing and information management, as well as an honors degree in computing with First Class Honors. He is also the recipient of many prestigious awards for academic excellence and outstanding academic performance. Mr. Marmaridis is an MCP and has been working in the IT sector for several years, gathering very substantial experience in systems administration and networking, information systems security, and high availability systems including Web farms and DB clusters. He is currently undertaking his PhD in computing and also enjoys golf, Toastmasters writing, and blogging on his Web site at <http://marmaridis.org/>.

N. Marsit is a PhD student at the Paul Sabatier University. His research work focuses on mobile database queries.

Israel Martin-Escalona earned a degree in telecommunications engineering from the Technical University of Catalonia (UPC) in 1999 and is currently an assistant professor at the Department of Telematic Engineering. Since 2001 he is involved in the IST Emily Project, funded by the European Commission and dealing with location services and technologies for GSM/GPRS networks. In 2002 he joined the project PPT, funded by Telefónica, on location services for the UMTS network. In 2004 he joined the IST Liaison Project on location services for WLAN and ad-hoc networks.

Henrique M. G. Martins is a medical doctor who also holds a master's in management Studies and is currently finishing his PhD at the University of Cambridge, working within the Information Management area on the topic: "The Use of Mobile Information and Communication Technologies in Clinical Settings." His interests are primarily in the interfaces between medicine and management

as facilitated by mobile clinical information systems, as well as management education to medical students and junior doctors. He has presented work at several conferences and has forthcoming articles published in these areas.

Carolyn McGregor holds a PhD in computing science, wherein she developed new ways to use intelligent decision support systems to assist organizations in the monitoring of business performance through the use of workflow audit logs. Furthermore, she has also provided strategic guidance in the areas of decision support, data warehousing, and data mining to some of Australia's leading corporations. In 2001, Dr. McGregor was the first Australian PhD student to be awarded a three-month IBM PhD Research Internship at the IBM TJ Watson Research Center in Yorktown, New York. Currently, she is active within the CASE Research Center at University of Western Sydney, where her research focus has been on development of an intelligent decision support system for business process performance management, known as the solution manager service, which she has refined and applied to problem domains such as health and medicine.

Shailendra Mishra earned his ME in computer science and engineering from MLNREC, (now MNIT Allahbad), Allahbad, and his master's degree in electronics in 1992 from the University of Allahbad, India. From August 1994 to February 2000, he led the Department of Computer Science and Electronics at ADC, University of Allahbad. From February 2000 to February 2001, he was with RG Engineering College, Meerut, affiliated with the UP Technical University, Lucknow, as assistant professor in the Department of Computer Science and Engineering. Presently, he is with the DehraDun Institute of Technology (DIT), affiliated with UP Technical University, as an assistant professor in the Computer Science & Engineering Department, and he is a joint faculty member of IGNOU, Delhi, India. His recent research has been in the field of mobile communication. He has also been conducting research on communication systems and networks with performance evaluation, and design of multiple access protocol for mobile communication networks. He is the author of more than 15 technical papers in international and national journals, and conference proceedings.

Manoj Misra received a BTech degree from HBTI Kanpur, India; an MTech degree from the Indian Institute of Technology, Roorkee; and a PhD from Newcastle upon Tyne, UK, in Computer Science & Engineering. He is currently an associate professor in the Electronics & Computer Engineering Department, Indian Institute of Technology, Roorkee. Dr. Misra's current research interests include mobile computing, performance evaluation, and distributed computing. He is also serving as co-coordinator of the Information Superhighway Center, IIT Roorkee.

Sulata Mitra received a BE in Electronics and Telecommunication and an ME in Power Electronics in 1986 and 1996, respectively. She is pursuing her PhD. Since 2000 she has been with the Department of Computer Science & Technology, Bengal Engineering and Science University (formerly Bengal Engineering College), where currently she is a lecturer. Her area of interest is mobile computing.

Guido Moerkotte studied computer science at the Universities of Dortmund, Massachusetts, and Karlsruhe from 1981 to 1987. The University of Karlsruhe awarded him a Diploma (1987), a doctorate (1989), and a postdoctoral lecture qualification (1994). In 1994 he became an associate

About the Authors

professor at RWTH Aachen. Since 1996 he holds a full professor position at the University of Mannheim, Germany, where he heads the database research group. His research interests include databases and their applications, query optimization, and XML databases. Dr. Moerkotte authored/co-authored more than 100 publications and three books.

F. Morvan is an associate professor at the Paul Sabatier University. His main research interests include parallel databases, distributed systems, and query response time optimization.

Manzur Murshed received his BEngg (*Honors*) in computer science and engineering from Bangladesh University of Engineering and Technology (BUET) in 1994, and his PhD in computer science from the Australian National University in 1999. He is currently the director of research and a senior lecturer at the Gippsland School of Computing and Information Technology, Monash University, Australia, where his major research interests are in the fields of multimedia communications, wireless communications, video coding and transcoding, video indexing and retrieval, video-on-demand, image processing, parallel and distributed computing, grid computing, simulation, complexity analysis, multilingual systems, algorithms, digital watermarking, and distributed coding. He has published more than 70 journal and peer-reviewed research publications and two refereed book chapters. He is the recipient of numerous academic awards including the *University Gold Medal* from BUET.

Sashi Nand holds BComm, MAcc, and PhD degrees, and provides administrative and accounting services to an engineering computer consultancy through Nandtech Pty Ltd. She has been a guest lecturer at the University of Technology Sydney Faculty of Engineering on the topic of Emerging New Technologies. Dr. Nand has published and presented papers at various international conferences including in Egypt and Perth. She is the winner of the PriceWaterhouseCoopers Prize in Advanced Accounting Theory, a member of the Australian Institute of Company Directors, as well as a member and ambassador for the Australian Computer Society. She was recently awarded as Toastmaster of the Year 2004/05 by Hornsby Toastmasters, and was designated Competent Leader and Competent Toastmaster.

Devon Nugent is a PhD candidate at the University of Queensland. Her research interests include Geographical Information Systems and mapping in location-based services.

V.S. Palsule received an ME in communication systems (Spread Spectrum Systems) from Jabalpur University, India, in 1979 and then joined the Space Applications Centre, ISRO, India. Presently he is heading the Advanced Communication Technology Division (ACTD) and is project director for EDUSAT, a satellite dedicated to education. He handled many prestigious projects of national interest, which includes spread spectrum systems for satellite communications, small communication terminal (SCOT), GPS, and MSS.

Amol Patel is founder and president of ConvergeLabs Corporation. Prior to that he was director of new ventures at ADC Telecommunications, a provider of integrated voice, video, and data solutions for the last mile of the communications network. His responsibilities included strategic planning, international mergers and acquisitions, and electronic commerce. Prior to ADC, he held

several positions in marketing and engineering at Intel Corporation, Cirrus Logic, and Sun Microsystems. He holds a BS (Honors) in Electrical Engineering and Computer Science from the University of California at Berkeley, an MS in Electrical Engineering from Stanford University, and an MBA from the Kellogg School of Management, Northwestern University.

Keyurkumar J. Patel received his First Class BEngg from the Bangalore University, India, and his MEngg (Robotics) from Swinburne University of Technology, Australia, in 1997 and 2000, respectively. He is currently pursuing his PhD. He is a Microsoft Certified Systems Engineer (MCSE), Microsoft Certified Trainer (MCT), Cisco Certified Academy Instructor (CCNA/CCAI), and Certified Novell Engineer (CNE), Novell Academic Instructor (NAI). He is currently employed with the Centre for Computer Technology and Super Cisco Academy Training Centre (SuperCATC) at the Box Hill Institute of TAFE, Australia, as a leader of ICT Higher Education Programs. He was a founder of the Emerging Communications Technologies (ECT) Research Cluster in his previous position at the University of Ballarat, Australia. His research interests include a good mixture of IT and engineering, including communication technologies, application implementation, image processing, systems design, health informatics, and advanced manufacturing techniques. He has successfully executed research and commercial projects including for major IT companies like Cisco and Microsoft. To date he has published more than 40 research studies in international refereed conferences and journals. He is an IASTED Technical Committee member for the term 2004-2007, a member of the Australian Computer Society (MACS), and a member of IEEE (Communication and Computer Societies).

Stephen Paull has more than 30 years of experience working in software development, project management, and the design, implementation, training, and support of manufacturing-based ERP systems. In addition, for the last 16 years, he has been a senior lecturer in the School of Information Systems at the Victoria University of Technology. He is currently involved in the teaching of the SAP software package as part of the curriculum for undergraduate and postgraduate courses, specifically in the areas of manufacturing, project management, and programming (ABAP).

Priyatamkumar received BE and MTech degrees in electronics and communication from the Karnataka University and National Institute of Technology, Karnataka, India, in 1989 and 2004, respectively. He joined the BVB College of Engineering and Technology, Hubli Karnataka, India, in 1989, where he was engaged in teaching antennas and advanced communication systems. He is currently a senior faculty member doing research in WCDMA and mobile communication. He is a member of the Institute of Electronics and a member of the Indian Society of Technical Teachers (MISTE). He can be reached via e-mail at priyatam@bvb.edu.

Jon Tong-Seng Quah is an associate professor of electrical and electronic engineering, Nanyang Technological University. He lectures in both undergrad as well as graduate courses such as *Software Development Methodology*, *Software Quality Assurance and Project Management*, *Object-Oriented System Analysis and Design*, and *Software Engineering*. His research interests include financial market modeling using neural networks, software reliability, e-commerce, as well as Web and WAP technologies and applications. Other than academic services, Dr. Quah has undertaken joint projects with major companies in banking and airline industries, as well as

About the Authors

statutory boards of the government body. Prior to his academic pursuit, Dr. Quah was a director of a local company dealing with industrial chemicals.

Mahesh S. Raisinghani is an associate professor at TWU School of Management's Executive MBA program. He is also a certified e-commerce consultant and a project management professional (PMP). Dr. Raisinghani was the recipient of the 1999 UD Presidential Award and the 2001 King Haggar Award for excellence in teaching, research and service. His previous publications have appeared in *Information and Management*, *Information Resources Management Journal*, *Journal of Global IT Management*, *Journal of E-Commerce Research*, *Information Strategy: An Executive's Journal*, *Journal of IT Theory and Applications*, *Enterprise Systems Journal*, *Journal of Computer Information Systems*, and *International Journal of Information Management*, among others. He serves as an associate editor and on the editorial review board of leading information systems/e-commerce journals and on the board of directors of Sequoia, Inc. Dr. Raisinghani is included in the millennium edition of *Who's Who in the World*, *Who's Who Among America's Teachers*, and *Who's Who in Information Technology*.

S. Rajeev is with the Department of Electronics & Communication Engineering, PSG College of Technology, Coimbatore, India. He has more than 13 years of industrial and academic experience. His research interests are policy provisioning systems and distributed computing and systems. He has authored three books in the area of computer communication.

K. R. Ramakrishnan received BE, ME, and PhD degrees in electrical engineering from the Indian Institute of Science, Bangalore, India, in 1974, 1976, and 1983, respectively. He is currently a professor with the Department of Electrical Engineering, Indian Institute of Science. His research interests include image processing, computer vision, medical imaging, water marking, and multimedia communication.

K. Ramamurthy is the Roger L. Fitzsimonds scholar and professor of MIS at the University of Wisconsin-Milwaukee. He has a bachelor's degree in mechanical engineering, a graduate diploma in Statistical Quality Control and Operations Research, and an MBA. He also earned his PhD in management information systems from the University of Pittsburgh. He has nearly 20 years of industry experience and has held several senior technical and executive positions prior to entering academia. His current research interests include electronic commerce including Internet; adoption, implementation, and diffusion of modern information technologies; supply chain management; strategic IS planning; self-directed teams; business process reengineering; and management of computer-integrated manufacturing technologies. He has published more than 30 articles in major scholarly journals like *MIS Quarterly* (where he is an associate editor) and *IEEE*. He is a charter member of the AIS, and was elected to the Beta Gamma and Sigma honor society.

Anne-Marie Ranft has a BS in computing science from the University of Technology, Sydney, and has seven years of commercial IT experience in Internet/e-business applications. She is currently completing her MS in computing. Her research interests include application of mobile technologies to globalization.

N. Raghavendra Rao is a professor at the SSN School of Management & Computer Applications, Madras. He holds a PhD in Finance from the University of Poona. His teaching experience in the disciplines of information technology and financial management spans 10 years. He has more than two decades of experience in the development of application software related to manufacturing, service-oriented organizations, financial institutions, and business enterprises. He regularly contributes IT articles to mainstream newspapers and journals.

Gina Reyes is a lecturer in the School of Information Systems at Victoria University. Her main teaching areas are in systems analysis, systems implementation, project management, and management of IT. She received her PhD from RMIT University with her thesis on the impact of organizational context on IS performance. Her current research interests include mobile computing applications and the effective design of e-learning systems.

Fabien Roméo is a graduate student in computer science (MSc). He is currently preparing his thesis on the administration of wireless software components.

A. F. Salam is an assistant professor in the ISOM Department at the University of North Carolina-Greensboro. He earned both his PhD and MBA degrees from the School of Management at SUNY Buffalo. His research has been published or is forthcoming in *Communications of the ACM*, *IEEE Transactions on Systems, Man and Cybernetics*, *Information & Management*, *Communications of the AIS*, *Information Systems Journal*, and *Information Systems Management*. He has also been a co-editor of a special issue of the *Journal of Electronic Commerce Research* and co-guest editor of special sections of the *CACM on Internet and Marketing*, and the *Semantic E-Business Vision*.

Jari Salo is a researcher with the Faculty of Economics and Business Administration, University of Oulu, Finland. His present research interests include digitization of business relationships and networks, electronic commerce, and mobile marketing.

Richard Schilhavy is a PhD student in the Department of Information Systems and Operations Management (ISOM) at the Bryan School of Business and Economics at the University of North Carolina-Greensboro. His research interests are in technology innovations and their influence on business functions and implications for supply/value chain operations. He has published papers in the proceedings of INFORMS and DSI conferences which cover issues related to queuing problems, information assurance, and software development.

B. Shankaranand received his BE in electronic and communication engineering from Mysore University, India, in 1973, and his MSc (Engg) in Microwave Engineering from the Kerala, India, in 1977. He joined the National Institute of Technology, Karnatak, India, in 1979, where he has been engaged in teaching and research and development of broadband communication systems, focusing on communication applications, especially in satellite, optical, and microwave applications. He is currently a professor at NITK India, and is a member of ISTE, IETE, and MIE. He can be reached via e-mail at bsnanda2000@yahoo.co.uk.

About the Authors

S. Shanmugavel graduated from the Madras Institute of Technology in Electronics and Communication Engineering in 1978. He obtained his PhD in the area of Coded Communication and Spread Spectrum Techniques from the Indian Institute of Technology (IIT), Kharagpur, in 1989. He joined the faculty of the Department of Electronics and Communication Engineering at IIT as a lecturer in 1987 and became an assistant professor in 1991. Presently, he is a professor in the Department of Electronics and Communication Engineering, College of Engineering, Anna University, Chennai, India. He has published more than 68 research papers in national and international conferences, and 15 research papers in journals. He was awarded the IETE-CDIL Award in September 2000 for his research efforts. His areas of interest include mobile ad hoc networks, ATM networks, and CDMA engineering.

Pramod Sharma is the director of the IT Program at the Cooperative Research Center for Sustainable Tourism and is based at the University of Queensland in Brisbane, Australia. His research interests include Geographical Information Systems, and the development of information and communications technology applications in tourism.

Khaled Shuaib holds a PhD in electrical engineering from the City University of New York (1999). While doing his graduate studies, he worked for two years as a senior member of the technical staff at GTE Labs (currently Verizon Technology Center) in Waltham, Massachusetts. In 1999, Dr. Shuaib joined Ascend Communications (currently Lucent Core Switching Division) in Westford, Massachusetts, as a principle performance engineer, where he worked on performance evaluation, reliability, and scalability of IP and ATM networks of ISPs. Since September 2002, he has been with the College of Information Technology at the UAE University as an assistant professor. Currently, he is the program coordinator of the Network Engineering Track and the director of the UAEU Cisco Regional Networking Academy. His research interests are in the areas of network design and performance, QoS IP networks, and protocols for ad-hoc and wireless networks.

Nipur Singh received her PhD from Gurukul Kangari University and her MCA from Banasthli Vidaypith, India. She headed the Department of computer science, K.G. second campus at Gurukul Kangari University. She has authored numerous technical papers in national and international journals and conference proceedings, which include her recent area of research in the field of mobile communication and mobile computing.

Rahul Singh is an assistant professor in the ISOM Department at the Bryan School of Business and Economics at the University of North Carolina-Greensboro. His research interests include intelligent multi-agent architectures, Semantic Web, emerging technologies, and agents in supply chain. His published research includes articles in the *Communications of the ACM*, *E-Services Journal*, and *International Journal of Production Economics*. He has also served as co-chair of the mini-track on Semantic E-Business at AMCIS 2004. He also co-guest edited the special section on “Semantic E-Business Vision” in the December 2005 issue of *Communications of the ACM*. He earned his PhD from Virginia Commonwealth University.

Jaakko Sinisalo is a researcher with the Faculty of Economics and Business Administration, University of Oulu, Finland. His research interests lie at the intersection of mobile marketing and customer relationship management.

S. N. Sivanandam heads the Department of Computer Science and Engineering, PSG College of Technology, Coimbatore, India. He has more than 25 years of professional teaching experience. His research interests include control systems, networking, distributed architectures, and soft computing. He has over 250 publications to his credit, and he has authored five books in the areas of control systems and soft computing.

K.V. Sreenaath is a student doing research in the Department of Information Technology, PSG College of Technology, Coimbatore, India. His research is in cryptography, service-level provisioning, and automated systems.

Hsiang-Ju Su is a graduate student from the Institute of Enterprise Innovation and Development, Department of Information Management, Shih Chien University. Su's main research activity addresses consumer behaviors in mobile phone services.

David Taniar received his bachelor's (Honors), master's, and PhD degrees in computer science, with a particular specialty in databases. His research now extends to data mining and mobile databases. He has published research papers in these fields extensively including a recent book *Object-Oriented Oracle*. Dr. Taniar now works at Monash University, Australia. He is an editor-in-chief of several international journals, including *Data Warehousing and Mining*, *Business Intelligence and Data Mining*, *Web Information Systems*, *Web and Grid Services*, *Mobile Multimedia*, and *Mobile Information Systems*. He is a fellow of the Institute for Management Information Systems.

Ramaprasad Unni is an assistant professor of marketing at Portland State University. His research interests are in understanding the role of technology (including the Internet) in marketing. He has also conducted several projects that have examined differences in consumer behavior in online and off-line environments, and has presented his research at several national and international conferences.

Khimji Vaghjiani has spent 16 years in the IT industry, covering manufacturing, telecommunications, and banking and finance. He has extensive experience advising organizations in the field of emerging technologies, with particular focus on the user needs with mobile technologies. He is currently working on his PhD at the University of Technology, Sydney, in the area of innovation and its application in management and transitions.

Ketan Vanjara is a lead program manager with Microsoft in its Global Delivery Center, India (GDCI). Previously, he held several senior positions in the software industry and has led large software enterprise software development teams on diverse domains and technology platforms. During his 17+ years in the software industry, he has worked with various delivery models on multiple platforms. He specializes in the creation of application software for various industry verticals like

About the Authors

health care, knowledge services, and manufacturing. He has in-depth knowledge and experience of working on various software development models, and implementing quality processes based on frameworks like CMM and ISO. He has published in numerous industrial as well as research journals and has contributed to chapters in edited books.

Amrish Vyas is in the process of completing his PhD at the University of Maryland-Baltimore County. Prior to joining the PhD program, he earned his MBA in Business Computer Information Systems from the Zarb School of Business, Hofstra University, USA, and his LLB from the Maharaja Sayajirao University, India. His research interests encompass expert (intelligent) systems, including intelligent agent systems and their application to an array of business problems, including location-aware applications in mobile services domain.

Fu Lee Wang received BEngg (Honors) in computer engineering and MPhil in computer science and information systems degrees from the University of Hong Kong, in 1997 and 1999, respectively. He earned his PhD in systems engineering and engineering management from the Chinese University of Hong Kong in 2003. He is currently an instructor in the Department of Computer Science at the City University of Hong Kong. He is a member of ACM and IEEE. His research interests include document summarization, digital library, and information retrieval.

Jia Jia Wang received an MSc with distinction in personal, mobile, and satellite communications from the University of Bradford, UK, in 2003, and a First Class BSc in electrical and electronic engineering from the HeiLongJiang University in China in 2002. In 2003, she joined the Mobile, Satellite, and Communication Research Center at the University of Bradford, where she participated in the EC's IST FIFTH project. She is currently in the second year of her PhD study; her research interests include mobile commerce, security, wireless mobility, and network protocol.

Tom Wiggen received his PhD from the Louisiana State University in 1973. He is an associate professor in the Department of Computer Science, University of North Dakota. His research interests include intelligent systems, modeling and simulation, and computer security.

David C. Wyld serves as the Mayfield professor of Management at Southeastern Louisiana University, where he also directs the Strategic E-Commerce Initiative. Dr. Wyld is a widely published author and invited presenter in the area of RFID technology. He is currently working on a report on RFID in the public sector for the IBM Center for the Business of Government and writing a book on RFID for Cambridge University Press.

Christopher C. Yang is an associate professor in the Department of Systems Engineering and Engineering Management at the Chinese University of Hong Kong. He received his BS, MS, and PhD degrees in electrical and computer engineering from the University of Arizona. Before he joined the Chinese University of Hong Kong, he was an assistant professor in the Department of Computer Science and Information Systems and the associate director of the Authorized Academic Java Campus at the University of Hong Kong. He has also been a research scientist in the Artificial Intelligence Laboratory in the Department of Management Information Systems at the University of Arizona. His recent research interests include cross-lingual information retrieval, multimedia

information retrieval, digital library, information visualization, Internet searching, automatic summarization, information behavior, and electronic commerce. He has published more than 100 refereed journal and conference papers including in the *Journal of the American Society for Information Science and Technology* and *IEEE*.

Hung-Jen Yang received his PhD from the Iowa State University in 1991. He is an associate professor in the Department of Industrial Technology Education, National Kaohsiung Normal University, Taiwan. His research interests include computer networks, automation, and technology education.

Victoria Yoon is an associate professor in the Department of Information Systems at the University of Maryland-Baltimore County. Before that, she was an associate professor in the Department of Information Systems at Virginia Commonwealth University. She received her MS from the University of Pittsburgh and her PhD from the University of Texas at Arlington. Her primary research interests are the application of expert systems to business decision making in organizations and the managerial issues of such systems.

Rupert Young is currently a reader at the University of Sussex; he graduated with a degree in Engineering from Glasgow University in 1984. Subsequently, he was employed as a research assistant and research fellow at Glasgow, during which time he gained wide experience in image and signal processing techniques. He participated in two Brite/EuRam collaborative European programs during that time, the second with Glasgow as the consortium leaders. His research led to the award of the PhD degree in 1994. In April 1995 he took up his present post as a lecturer in the School of Engineering, University of Sussex. He was promoted to senior lecturer in 1997 and then reader in 1999. He has disseminated his work to the research community by publication of more than 80 refereed journal and conference papers. He is a member of the Institute of Electronic and Electrical Engineers.

Symbols

2.5G Technology 146
 2G (see second generation)
 3G (see third generation)
 3GPP (see third-generation partnership project)
 4G (see fourth generation)
 8-PSK (8-Phase Shift Keying) 147

A

a posteriori probabilities (APPs) 188
 a priori 188
 abstract data type 275
 model 276
 abstract location 60
 ACA (see adaptive conjoint analysis)
 access cost 615
 accessibility 697
 accuracy 39
 ad hoc 308
 adaptation 609
 adaptive 8
 conjoint analysis (ACA) 759

adaptive modulation and coding (AMC) 152, 153
 additive white Gaussian noise (AWGN) 186
 administrative operations 769
 advanced
 communication age 719
 encryption standard (AES) 396, 496
 medical priority dispatch 106
 mobile phone system (AMPS) 144
 advertising 607, 608
 AES (see advanced encryption standard)
 affordability 505
 after-purchase evaluation 634
 agent 3, 367
 activation 373
 freeze 371, 373
 ontology 511
 pre-activation 372
 receptionist 370
 transport 372, 373
 agent's
 fitness function 532
 GUI 534

knowledge 533
 lifecycle 532
 task 533
 AGPS (see assisted GPS)
 AIDC (see automatic identification and data capture)
 alignment with business strategy 599
 always-on capability 467
 AMC (see adaptive modulation and coding)
 AMPS (see advanced mobile phone system)
 analog technology 143
 angle of arrival (AOA) 3
 anonymous 7
 anti-counterfeiting technologies 381
 AP mobile time and travel 722
 aperiodic 554
 APIs/SDKs 51
 APPs (see a posteriori probabilities)
 application 431
 in marketing 223
 programs 409
 architecture 515
 design 557
 ARPU (see average revenue per user)
 ASICs 153
 assistance data provider 51
 assisted GPS (AGPS) 45, 507
 asynchronous 7, 133, 174
 invalidation report 133
 authentication 396, 412, 617
 technologies 381
 authorization 412
 automatic
 identification and data capture (AIDC) 382
 summarization 250
 vehicle location system 106
 autonomous 8
 autonomy 682
 availability 412, 505
 average revenue per user (ARPU) 19
 AWGN (see additive white Gaussian noise)
 ayurveda 116

B

B2B (see business-to-business)
 B2C (see business-to-consumer)
 B2E (business-to-employee)
 BAA (see bandwidth adaptation algorithm)
 bandwidth 803
 adaptation algorithm (BAA) 357, 359

constraints 620
 cost 615
 units (BUs) 351
 banking industry 765
 barcode 383
 base station (BS) 135, 174, 297
 basic
 complex scrambling 168
 formats of mobile education 497
 batteries 411
 battery power constraints 618
 BBB (see Better Business Bureau)
 behaviourgraphic 648
 benchmarking 421
 benefits 475
 BER (see bit error rate)
 Better Business Bureau (BBB) 458
 billing sscheme 701
 bioinformatics 114, 511
 bit error rate (BER) 153, 186
 blood glucose monitor systems/diabetes systems 119
 Bluetooth 98, 152, 184, 323, 445, 497, 556, 585, 805
 technology 3, 21
 branch applications 773
 branding 531
 broadband connectivity 185
 BS (see base station)
 BUs (bandwidth units)
 budget spend 465
 business
 anywhere 464
 goals 592
 intelligence 118
 models 432
 process 829
 reengineering 720
 business-to-business (B2B) 445, 676
 business-to-consumer (B2C) 445
 business-to-employee (B2E) 479

C

C2C (see consumer-to-consumer)
 CA (see certification authority)
 CAC (see call admission control)
 cache
 invalidation strategy 133, 134
 state information (CSI) 134, 136
 caching 133

Index

- CAD (see computer-aided dispatch)
- call
 - admission control (CAC) 335, 352
 - blocking probability (CBP) 334
 - dropping probability (CDP) 334
 - drops 609
- capability 505
- capital expenditure 769
- care provider 84
- CBP (see call blocking probability)
- CDMA (see code division multiple access)
- CDP (see call dropping probability)
- CDPD (see cellular digital packet data)
- cell ID (see cell identification)
- cell
 - identification (cell ID) 40
 - search
 - algorithm 177
 - design 176, 178
 - visiting
 - probability (CVP) 339
- cellular digital packet data (CDPD) 406
- Cerner Pocket Power Chart (CPPC) 99, 102
- certificates 102
- certification authority (CA) 396
- channel
 - coding 154
 - estimation 187
- characteristics 70
- children tracking 36
- chipping rate 146
- chips 146
- Citrix 101
- client 133
 - device flexibility 505
 - pull 554
- clinical
 - alerts 86
 - management 84
 - work practices 69
- clusters integration 605
- code
 - division multiple access (CDMA) 2, 144, 158, 488, 585
 - group identification 177
- collaborative 8
- commission reporting and tracking 774
- common
 - pilot channel (CPICH) 175
 - standards 128
- CMMS (see computerized maintenance management system)
- communication 607, 608
- communicative 9
- commuting habits 699
- compact pattern (CP) 298
- company intranet 774
- competitive and product information delivery 774
- complex scrambling 160
- complexity 115
- compression ratio 254
- computer
 - aided dispatch (CAD) 106
 - assisted passenger profiling system 746
- computerized
 - reservation systems (CRSs) 810
 - maintenance management system (CMMS) 733
- conceptualization 55
- confidentiality 396, 412
- connected 505
 - mobile tourist 814
- connection stability 803
- connectivity 609
- consistency 40
- constraint database model 275
- constructing questionnaires 635
- consulting physician station (CPS) 87
- consumer-to-consumer (C2C) 445
- content 28
 - delivery 695
 - management agent 12
- context 2
- continuous query (CQ) 6, 270
- convergence 116
- COPS-PR (see COPS for provisioning)
- COPS for provisioning (COPS-PR) 618
- cost 609
 - paradox 118
 - price paradox 115
- counterfeiting 380
- coverage 609
- covert technologies 381
- CP (see compact pattern)
- CPICH (see common pilot channel)
- CPPC (see Cerner Pocket Power Chart)
- CQ (continuous query)
- CRC (see cyclic redundancy check)
- CRM (see customer relationship management)
- crossover 519

CRSs (see computerized reservation systems)
 CSD 176
 CSI (see cache state information)
 current 2
 custom declaration service 500
 customer
 differentiation 661
 needs 424
 relationship management (CRM) 585, 605, 643, 767
 saturation 702
 -centric value 424
 CVP (see cell visiting probability)
 cyclic redundancy check (CRC) 388

D

DAML-OIL 55
 DAML-S 63
 DAP (see directory access protocol)
 data
 collection unit (DCU) 87
 delivery mechanism 554
 driven 116
 integrity 396
 mining 217
 transmission speed 609
 warehouse 831
 database servers 408
 DCU (see data collection unit)
 decision-feedback equalizer (DFE) 187
 dedicated physical data channel (DPDCH) 159
 defuzzification 311
 delayed uplink (DU) 138
 delivering the position 49
 demand-responsive transport 36
 democratisation 794
 demographics 656
 denial-of-service (DoS) 621
 destination management system (DMS) 813
 developing nations 656
 DFE (see decision-feedback equalizer)
 diagnostics 114, 118
 differential GPS 45
 differentiated services (Diffserv) 619
 Diffserv (see differentiated services)
 Diffserv-based wireless network 619
 digital
 divide 656
 purchases 471
 subscriber line (DSL) 487

direct sequence (DS) 158
 spread spectrum (DSSS) 324
 directory access protocol (DAP) 620
 disconnection time of a client 133
 discussion 636
 distance smoothing factor (DSF) 344
 distribution 864
 distributor 389
 DMS (see destination management system)
 doctors 69
 DPDCH (see dedicated physical data channel)
 DS (see direct sequence)
 DSF (see distance smoothing factor)
 DSL (see digital subscriber line)
 DSL and security 98
 DSPs 153
 DSSS (see direct sequence spread spectrum)
 DU (see delayed uplink)
 dynamic 7
 source routing (DSR) 309
 topological changes 620

E

EAF (see exponential averaging function)
 EAP (see extensible authentication protocol)
 earliest deadline first (EDF) 311
 early
 adopters 646
 majority 646
 -stage identification 756
 e-business (see electronic business) 507
 ECC (error control code)
 e-customer (see electronic customer) 656
 EDF (see earliest deadline first)
 EDGE (see enhanced data rates for GSM evolution)
 education 607
 efficiency gains 770
 electromagnetic
 (EM) waves 143
 pulses (EMPs) 874
 electronic
 commerce (e-commerce) 435, 463, 483
 medical record (EMR) 74, 126
 product code (EPC) 382, 861
 e-mail 607, 773
 EMPs (see electromagnetic pulses)
 EMR (see electronic medical record)

Index

- encryption
 - and secure data transfer 771
 - toggling 772
- energy constrained nodes 621
- enhanced
 - cell ID (Cell ID++) 41
 - data rates for GSM Evolution (EDGE) 147
 - observed time difference (E-OTD) 42
 - services 467
- enterprise resource planning (ERP) 605, 697, 840, 870
- entertainment 695
- E-OTD (see enhanced observed time difference)
- EPC (see electronic product code)
- ERP (see enterprise resource planning)
- error
 - control code (ECC) 188
 - resilient coding 154
- ETD (see expected travelling distance)
- e-transformations 563
- European Universal Mobile Telephone System (UMTS) 173
- evaluation
 - of acquired information 634
 - research method 100
- evolution
 - of PDA devices 97
 - process 536
- executive dashboards 774
- expected travelling distance (ETD) 344
- expert system 608
- exploitation of ECC 187
- exponential averaging function (EAF) 344
- extended enterprise 473
- extensibility 571
- extensible
 - authentication protocol (EAP) 102, 397
 - Markup Language (XML) 86, 545
 - style sheet transformation) 545

F

- fair borrowable bandwidth (FBB) 354
- fairness 685
- Fanyu Project 499
- FBB (see fair borrowable bandwidth)
- FEC (see forward error correction)
- feedback 519
- FHs (see fixed hosts)
- field service 477
- FIMP (see first informer mobile phone)

- financial markets 781
- fingerprint 102
- fingerprinting 46
- finite impulse response (FIR) 191
- FIR (see finite impulse response)
- firmographics 661
- first
 - generation (1G) wireless technologies 144
 - informer mobile phone (FIMP) 107
- fitness function 513, 515
- fixed
 - hosts (FHs) 134
 - wireless networks 406
- flexible 8
- forensic technologies 382
- forward error correction (FEC) 144
- fourth generation (4G) 551, 724
 - mobile systems 184
 - wireless technologies 152
- FPGAs 153
- fractal summarization 253
 - model 249, 252
- frame synchronization 177
- frequency
 - division
 - duplex (FDD) 174
 - multiple access (FDMA) 144
 - hopping (FH) 158
 - modulation (FM) 144
- full
 - duplex 144
 - function device (FFD) 325
- fully
 - adaptive 334
 - elastic 334
- function-specific handheld or nano devices 117
- functionalities 6, 48
- functionality description 64
- fuzzification 311, 313

G

- gambling 447
- gateway mobile location centers (GMLCs) 25, 48
- gateway MLC (see gateway mobile location centers)
- gathering the location requests from several GMLCs 48
- Gaussian minimum shift keying (GMSK) 147
- GDSs (see global distribution systems)

general
 agent transport 368
 message format 369
 packet radio service (GPRS) 147, 488, 551, 585, 631, 671, 768
 generations 143
 genetic algorithm 512
 geographic
 information system (GIS) 793, 843
 positioning systems (GPS) 864
 GI monitoring system 120
 GIS (see geographic information system)
 global
 differences 656
 distribution systems (GDSs) 810
 positioning system (GPS) 2, 21, 45, 114, 174, 240, 507, 659, 667, 696, 846
 roaming 149, 154
 system for mobile communication (GSM) 144, 334, 470, 489, 551, 585, 832
 GMLC (see gateway mobile location centers)
 goal-oriented 8
 Golay correlator 176
 government applications 607
 GPRS (see general packet radio service)
 GPS (see global positioning system)
 graph-based ontology 59
 Green Pages 65
 GSM (see global system for mobile communication)

H

handheld
 device 422, 466, 589
 MRI 120
 PDAs 474
 handoff techniques 154
 handset
 module 652
 -based technologies 21
 HDTV 155
 heading feature in fractal summarization 257
 HEAL (see hospital and emergency ambulance link)
 health hazards 609
 healthcare 114
 administration 114, 118
 information management system (HIMS) 118, 126

hierarchical display 248
 high
 network capacity 185
 speed
 circuit-switched data (HSCSD) 147
 downlink packet access (HSDPA) 152
 HIMS (see healthcare information management system)
 HiperLAN 445
 holistic medicine 121
 home-working 478
 homeopathy 116
 hospital 69
 and emergency ambulance link (HEAL) 107
 host computers 407, 415
 hotspots 333, 441
 HotSync 99, 100
 HSCSD (see high speed circuit-switched data)
 HSDPA (see high speed downlink packet access)
 HTML (see HyperText Markup Language)
 HTTP protocol 555
 human resources 606
 hybrid
 techniques 46
 technologies 21
 HyperText Markup Language (HTML) 545
 hypotheses 634

I

ICT (see information and communication technology)
 ICU (intensive care unit)
 Id-Synch and P-Synch 616
 identity management 507
 idle period downlink (IPDL) 43
 improved mobile telephone service (IMTS) 143
 i-mode 405
 IMT-2000 (see International Mobile Telephone 2000)
 IMTS (see improved mobile telephone service)
 increased
 bandwidth 467
 data encryption 396
 personal productivity of employees 587
 incremental redundancy 148
 individual privacy 668
 indoor GPS 45
 industrial automotive sector 475

Index

- inferential capability 9
 - information and communication technology (ICT)
 - 142, 564, 602, 719
 - information
 - gathering 634
 - retrieval 512
 - services 36
 - on the e-tap 471
 - stored in MSC 303
 - system (IS) 83, 676
 - technologies (ITs) 83, 675
 - InfoSpiders 530
 - infrared (IR) 92
 - innovation 686
 - innovators 646
 - input/output devices 411
 - institutional banking 781
 - insurance 476
 - integration of devices, applications, and platform 127
 - integrity 412, 609, 780
 - intelligent
 - advertising 447
 - agents 7
 - install and update 505
 - migration 290
 - intensive care unit (ICU) 83
 - inter-operability between carriers 467
 - interactive selling tools 775
 - interim standard
 - 136 (IS-136) 145
 - 95 Code Division Multiple Access 145
 - intermittent communication links 556
 - International
 - Mobile Telephone 2000 (IMT-2000) 149
 - Telecommunications Union (ITU) 149
 - Internet 491, 709
 - service providers (ISP) 487, 555
 - access-based mobile education 497
 - enabled mobile handheld devices 402, 403, 409, 414, 415
 - interoperability 23, 31, 48, 107, 242, 425, 440, 598
 - invalidation report (IR) 133
 - investments and infrastructure 127
 - IPDL (see idle period downlink)
 - iPod 119
 - IR (see infrared)
 - IS-136 (see interim standard 136)
 - IS (see information system)
 - ISI (inter symbol interference) 153
 - iSilo 99
 - ISP (see Internet service provider)
 - issues and challenges 479
 - ITs (see information technologies)
 - ITU (see International Telecommunications Union)
- ## **J**
- Java
 - frameworks 570
 - management extensions 203
 - banking 425
 - joint source channel coding 154
- ## **K**
- keyboards 411
 - keyword queries 514
 - knowledge
 - management 606
 - repository 831
 - based 656
 - system 830
- ## **L**
- laggards 646
 - LAN (see local area network)
 - landscape 114
 - language 128
 - laptop 70
 - computers 218
 - LAQ (see location-aware query)
 - large-scale distributed parallel mobile (LDPM) 499
 - late majority 646
 - latency 803
 - laws 128
 - LBS (see location-based service)
 - manager 51
 - LDAP access engine and directory structures 616
 - LDPC (see low density parity check)
 - LDPM (see large-scale distributed parallel mobile)
 - LDQ (see location-dependent query)
 - leisure 607
 - LIA (see location-aware intelligent agent)
 - LIF (see location interoperability forum)
 - lifestyle 430

- limited
 - memory 615
 - security 621
- LMU (see location measurement unit)
- local
 - area network (LAN) 406, 432
 - multipoint distribution service (LMDS) 407
 - server-based mobile 497
- localization 697
 - of mobile units 272
- location
 - client (LCS client) 48
 - feature in fractal summarization 256
 - interoperability forum (LIF) 819
 - management agent 11
 - measurement unit (LMU) 43, 49
 - middleware 51
 - models 272
 - ontology 58
 - pattern matching (LPM) 3
 - server provider cluster 25
 - service 273
 - services 35
 - system architectures 47
 - aware
 - computing 2
 - intelligent agent (LIA) 3, 10
 - query (LAQ) 269
 - based service (LBS) 18, 36, 447, 472, 507, 798
 - dependent query (LDQ) 269
 - sensing technologies 6
- logistic company 389
- logistics 475
- lone worker applications 36
- loose coupling 46
- lost work 609
- low
 - bandwidth 618
 - computational complexity 187
 - density parity check (LDPC) 154
 - codes 188
- LPM (see location pattern matching)
- M**
- MAC (see media access control)
- macro
 - economic level 451
 - payment and micro-payment methods 413
 - payments 404, 413
- MAI (see multiple access interference) 153
- management
 - barriers 770
 - layer 11
- managing
 - location
 - requests 48
 - responses 48
 - mobile environment 288
 - technology 609
- manufacturers 863
- market analysis 496
- marketability 659
- marketing 223, 431, 454
 - issues 609
- maximise ROI 599
- maximum-likelihood (ML) 186
- m-banking (see mobile banking)
- mBPR (see m-business process engineering)
- MBPS (see measured based priority-queue scheme)
- m-business (see mobile business)
- MC-CDMA (see multi-carrier CDMA)
- m-commerce (see mobile commerce)
- MCS (see multiple modulation and coding schemes)
- mean squared error (MSE) 190
- means-end chain (MEC) 756
- measured based priority-queue scheme (MBPS) 337
- MEC (see means-end chain)
- media access control (MAC) 731
 - filtering 395
- medical
 - alert 36
 - calculator 99
 - information 126
 - programs on PDA devices 99
 - software programs 99
- m-education (see mobile education)
- memory 411, 608
- message
 - delivery mechanism 558
 - exchange 305
 - integrity check (MIC) 396
- messaging module 652
- meta directories 616
- method outline 100
- methodology perspective 789
- MGIS (see mobile GIS)
- MHs (see mobile hosts)

Index

- MHSs (see mobile healthcare systems)
- MIC (see message integrity check)
- micro
 - browsers 410
 - payment 404, 413, 447
 - surgeries 114
- middleware 286
- MIDP (see mobile information device profile)
- migration of nodes 289
- MIMS (see monthly index of medical specialities)
- ML (see maximum-likelihood)
- m-marketing (see mobile marketing)
- m-newspaper (see mobile newspaper)
- mobile
 - advertising 472
 - assisted techniques 39
 - banking 421, 471
 - applications 424
 - based techniques 39
 - business 393, 719, 721, 724
 - environment 444
 - process 586
 - engineering (mBPR) 720
 - services 632
 - central processing units 410
 - client 268
 - hardware providers 24
 - commerce 26, 55, 233, 366, 435, 436, 445, 463, 465, 694, 510
 - applications 404, 414, 540
 - ecosystem 430
 - payment methods 404, 412
 - revolution 528
 - system 402, 404
 - communications 773
 - computing 475, 613, 832, 841
 - devices 840
 - in healthcare 70
 - connectivity 599
 - consumer behavior 633
 - CRM (see customer relationship management)
 - customer 23
 - relationship management (m-CRM) 644, 712
 - devices 247, 285, 468, 723
 - market 722
 - e-commerce 445
 - mobile education 496
 - enabled business process 586
 - enterprise 587
 - systems 588
 - environment 288
 - equipment 218
 - GIS (MGIS) 794
 - government 499
 - hardware technology 132
 - healthcare systems (MHSs) 84
 - hosts (MHs) 134
 - information
 - and communication technologies 69
 - information device profile 202
 - infrastructure 599
 - Internet 422, 469
 - banking 607
 - module 652
 - shopping 607
 - location protocol (MLP) 48, 804, 819
 - maintenance 733
 - marketing 710
 - middleware 405, 414
 - network operator 24, 697
 - networks 614
 - mobile newspaper 608
 - nodes (MNs) 308
 - originating location request (MO-LR) 49
 - operating systems 409
 - payment methods 402
 - phone 491
 - technologies 632
 - phones 218, 425, 545
 - portals 24
 - queries scheduling 280
 - server 268
 - services 2
 - shopping 472
 - station (MS) 174
 - support stations (MSSs) 134
 - switching centre (MSC) 297, 333
 - technologies 494
 - technologies 563
 - technologies 114
 - unit (MU) 333
 - user 268, 298
 - data mining 217
 - types 472
 - VoIP 436
 - Web Services (MWSs) 545, 549
 - challenges 549
 - technology 107
 - workers 770

mobility 9, 114, 444, 770, 832
 and remote assistance 115
 model based on dynamic attributes 276
 modeling of moving objects 274
 moderate resolution 98
 MIMO (see multiple-input multiple-output)
 MMS (see multimedia message service)
 MO-LR (see mobile -originating location request)
 MODQs (see mobile object database queries)
 monitoring tools 535
 monthly index of medical specialities 99
 mortgage origination 775
 motivation issues 771
 moving
 object 268
 object database queries (MODQs) 270
 MS (see mobile station)
 MSC (see mobile switching centre)
 MSE (see mean squared error)
 MSSs (see mobile support stations)
 m-transformation 564
 m-transition 788
 MU (see mobile unit)
 multi-carrier CDMA (MC-CDMA) 153
 multi-hop communications 621
 multicast group size 319
 multicasting 7
 multimedia 148
 multimedia message service (MMS) 154, 490, 644, 669
 multimodality 506
 multiple
 authentication modes 772
 bearers 556
 -input multiple-output (MIMO) 152
 modulation and coding schemes (MCS) 148
 platforms 557
 multiplicity 115
 mutation 519
 MWSs (see mobile Web services)

N

nanotechnology 116
 natural interfaces 117
 naturopathy 116
 navigation 36
 information 51
 NCL (see Network Command Language)
 neonatal intensive care units (NICUs) 83

network
 -based positioning (or remote positioning) 711
 -based techniques 39
 -based technologies 21
 Command Language (NCL) 286
 drivers 51
 -induced location request (NI-LR) 49
 measurement report (NMR) 41
 security 146, 154
 neural network-based agent 290
 NICUs (see neonatal intensive care units)
 NI-LR (see network-induced location request)
 NLRQ (see non-location-related query)
 NMR (see network measurement report)
 NMT (see Nordic Mobile Telephone)
 node migration 290
 nomad 473
 non-elastic 334
 non-line-of-sight coverage 45
 non-location-related query (NLRQ) 268, 269
 non-medical information 127
 non-repudiation 396, 412
 non-synchronized coupling 47
 Nordic Mobile Telephone (NMT) 144

O

object name server (ONS) 861
 observed time difference of arrival (OTDOA) 43
 OFDM (see orthogonal frequency division multiplexing)
 off-line capable 505
 online journal 668
 ONS (see object name server)
 ontology 55, 531
 operations 769
 opt-in approach 448
 OQPSK (offset quadrature phase shift keying) 162
 organisational knowledge 568
 orthogonal
 complex 162
 quadrature phase shift keying 162
 variable spreading factor (OVSF) 175
 frequency division multiplexing 153
 OTDOA (see observed time difference of arrival)
 overt technologies 381
 OWL (see Web Ontology Language)
 OWL-S 63, 243
 oximeter 119

Index

P

- P/S management agent 11
- Pacific Digital Cellular(PDC) 145
- packet
 - delivery ratio (PDR) 319
 - switched networks 185, 467
- Palm 101
- PAN (see personal area network)
- paraplegia 114
- partner relationship management (PRM) 644
- Passport and Visa Enquiry Service 500
- pay-
 - as-you-go/prepaid mobile phones 454
 - per-visit agreements 235
- payment 607
 - mechanisms 433
- Payroll 606
- PCF (see position calculation function)
- PDA (see personal digital assistant)
- PDC (see Pacific Digital Cellular)
- PDP (see policy decision point)
- PDR (see packet delivery ratio)
- PEOU (see perceived ease of use)
- perceived
 - ease of use (PEOU) 677
 - usefulness (PU) 677
- performance 780
 - evaluation 314
- periodic 554
- personal
 - area network (PAN) 152, 325, 724
 - digital assistant (PDA) 70, 84, 96, 200, 218, 234, 248, 333, 410, 425, 435, 528, 545, 583, 655, 709, 733, 766, 779, 841
 - handphone systems (PHSs) 488
 - identification number (PIN) 425
 - information management (PIM) 584
- personalization 633, 645, 698
- personalized/customized 8
- pervasive information access 464
- PEs (see positioning elements)
- PFL (see predicted future location)
- pharmacogenomics 114
- PHSs (see personal handphone systems)
- Physical Markup Language (PML) 861
- physiological log Web service 88
- PIM (see personal information management)
- PIN (see personal identification number)
- PKI (see public key infrastructure)
- PML (see Physical Markup Language)
- PN (see pseudo-noise)
- pocket PC 101
- POI (see point of interest)
- point of interest (POI) 799
- points of presence (PoPs) 474
- policy
 - based architecture for QoS 617
 - computing 613
 - decision point (PDP) 619
 - warehouse 616
 - based
 - approach 614
 - architecture 615, 620
 - architecture for security 615
 - QoS 618
- PoPs (see points of presence)
- portable patient monitors 120
- portrait and landscape 98
- position calculation function (PCF) 39
- positioning elements (PEs) 44
- post-roaming notification 374
- pragmatics 397
- pre-roaming notification 373
- predictable availability 803
- predicted future location (PFL) 12
- presence awareness 507
- price 531
- pricing 29, 701
- primary synchronization channel (P-SCH) 175
- priority perceptions 465
- privacy 230, 609
 - and security 128
- privacy, spam, and security 645
- process efficiency 587
- processing of location-related operators 274
- procurement 606
- product
 - brokering agent 528, 531
 - matching 647
 - ontology 531
 - recommendation 533, 537
- profile 531
 - attributes 64
 - management agent 12
- profiling tools 775
- project
 - objectives 592
 - plan 592
 - progress 851
- protected extensible authentication protocol (PEAP) 397

- protection of privacy 230
- proteomics 114
- prototype
 - design and implementation 515
 - testing and evaluation 521
- provider information 64
- providing the service 617
- proxy-based mobile Web services 550
- P-SCH (see primary synchronization channel)
- pseudo-noise (PN) 160
- PSTN (see public-switched telephone network)
- PU (see perceived usefulness)
- public Internet connectivity 736
- public key infrastructure (PKI) 387, 396
- public-switched telephone network (PSTN) 143, 333
- publish/subscribe
 - middleware 11
 - model 558
 - paradigm 3
- publishers 10
- pull model 3
 - strategy 14
- pulse meter for Ayurveda 121
- purchasing 634
- push
 - model 3
 - strategy 13
 - technology 555

Q

- QoS (see quality of service)
- QPSK (see quadrature phase shift keying)
- quadrature phase shift keying (QPSK) 175
- quality of service (QoS) 37, 185, 310, 332, 362, 421, 614
- query
 - formation 511
 - restructuring 514
- querying
 - fixed-object databases 272
 - in
 - constraint database model 277
 - MOST model 277
 - moving object databases 276
- quick
 - cure 115
 - response 115
- quoting for insurance agents 774

R

- RA (see registration authority)
- radial span 350
- radio
 - frequency (RF) 589, 873
 - identification (RFID) 381, 589, 744, 859, 871
 - applications 861
 - network 861
 - transmission technology (RTT) 150
- range and average time 103
- RAP (see resource allocation protocol)
- RASs (see remote access servers)
- RDF (see resource description framework)
- reachability 697
- reactive 8
- recognition 685
- reduced function device (RFD) 325
- registration authority (RA) 396
- regulations 609
- relevance feedback 512
- remote
 - access servers (RASs) 474
 - evaluation (REV) 286
 - positioning (see network-based positioning)
 - procedure calls (RPCs) 286
 - supervision 608
- request
 - for
 - entry permit 373
 - roaming permit 371
 - unfreeze key 373
 - and agent activation 372
 - reports (RRs) 133
- requirement confirmation 633
- research
 - design 635
 - methodology 842
 - model 681
- reservation
 - protocol (RSVP) 619
 - time window (RTW) 346
- resource
 - allocation protocol (RAP) 618
 - constraint 115, 556
 - description framework (RDF) 55
 - reservation policies 351
- response time 45
- retailer 389

Index

- return
 - of alternative therapies 116
 - on investment (ROI) 574, 741, 769
 - returning and repurchasing 634
 - REV (see remote evaluation)
 - reverse
 - fundamental channel (R-FCH) 159
 - link (uplink) 160
 - pilot (R-Pilot) 159
 - supplemental channels (RSCHs) 159
 - rewards 449
 - RF (see radio frequency)
 - R-FCH (see reverse fundamental channel)
 - RFD (see reduced function device)
 - RFID (see radio frequency identification)
 - risk management 775
 - roaming user 473
 - ROI (see return on investment)
 - RPCs (see remote procedure calls)
 - R-Pilot (see reverse pilot)
 - RRs (see request reports)
 - RSCHs (see reverse supplemental channels)
 - RSVP (see reservation protocol)
 - RTT (see radio transmission technology)
 - RTW (see reservation time window)
 - rule evaluation 314
- S**
- SACL wireless authentication 731
 - sales
 - analysis and planning 775
 - force automation (SFA) 774
 - opportunity 477
 - SAP mobile
 - asset management 723
 - procurement 723
 - sales 722
 - service 722
 - SCA (see super composition agent)
 - scalability 133
 - requirements 615
 - scheduler performance 318
 - SCM (see supply chain management)
 - scrambling code identification 177
 - SDH (see service description header)
 - second-generation (2G)
 - networks 185
 - wireless technologies 144
 - secondary synchronization channel (S-SCH) 175
 - secure sockets layer (SSL) 772
 - security 394, 432, 609, 834
 - features 387
 - issues 474
 - selection of survival 519
 - self
 - confidence 758
 - starting/proactive 8
 - selling change management 605
 - semantics 397
 - sense of urgency 466
 - SER (see symbol-error-rate)
 - server
 - push 554
 - client architecture 107
 - service 56, 234
 - broker 56
 - composition 243
 - description header (SDH) 289
 - dictionary 56
 - grounding 64
 - personalization 185
 - profile 64
 - provider 599, 616
 - requester 56
 - set identifiers (SSIDs) 732
 - oriented architectures (SOAs) 54
 - SFA (see sales force automation)
 - SGML (see Standard Generalized Markup Language)
 - shadow cluster 335, 349
 - short
 - message service (SMS) 146, 154, 424, 459, 490, 584, 644, 661, 709
 - displays 449
 - banking 424
 - based mobile education 497
 - message system (SMS) 669
 - range technologies 21
 - signal-to-noise ratio (SNR) 42, 186
 - significant development 18
 - SIM (see subscriber identity module)
 - simple object access protocol (SOAP) 548
 - simulation
 - environment and methodology 315
 - results 306
 - single channel 305
 - slot synchronization 178
 - smart
 - cellular phones 70, 410
 - client 505

- cards 102
- SMLC (Serving MLC) 48
- snapshot queries 6
- sniffer solutions 772
- soap-RPC 548
- social
 - integration 759
 - interaction 128
- socio-cultural 694
- sociology perspective 789
- SOE (see standard operating environment)
- soft handoff 146
- software
 - blanking (SWB) 44
 - defined radio 154
 - infrastructure 468
 - maintenance 774
 - market 722
 - power 116
 - tools 831
- solution manager service (SMS) 88
- source coding 154
- spatial and temporal dimensions 70
- spatio-temporal queries 270
- speed sensitivity factor 345
- spread spectrum (SS) 158
- SRWM (see symmetric random walk model)
- S-SCH (see secondary synchronization channel)
- SSIDs (see service set identifiers)
- SSL (see secure sockets layer)
- staff productivity 770
- stakeholders 576
- standard
 - Generalized Markup Language (SGML) 545
 - operating environment (SOE) 568
 - Query Languages 280
- standardization 469, 609
- standards compliance 571
- stateless 7
- statistical analysis 100
- stem cell research 114
- stretch display 98
- structure of wireless networks 406
- subscriber 10
 - identity module (SIM) 111
- super composition agent (SCA) 243
- supervised agent transport 369
- suppliers 862
- supply chain management (SCM) 585, 605, 870

- support
 - services 769
 - software 409
- strengths, weaknesses, opportunities, and threats (SWOT) 451
- SWOT (see strengths, weaknesses, opportunities, and threats)
- symbol-error-rate (SER) 190
- symmetric random walk model (SRWM) 337
- synchronization 411
 - of patient information 100
- synchronized coupling 47
- synchronous 174
- Syntactics 397
- system evaluation 537
- systems security 600

T

- tablet PCs 70
- TACS (see total access communications system)
- TA-IPDL (see time-aligned idle period downlink)
- TAM (see technology acceptance model)
- taxation payment 499
- TCO (see total cost of ownership)
- TCP/IP 550
- TDD (time division duplex)
- TDL (see Template Description Language)
- TDMA (see time division multiple access)
- TDOA (time difference of arrival)
- teamwork 780
- technology 20, 432
 - acceptance model (TAM) 676, 677
 - constraints 755
 - domain 4
 - perspective 788
- telematic unit (TU) 872
- telemedicine 114
- telemetric 607
- Template Description Language (TDL) 615
- temporal key integrity protocol (TKIP) 396, 731
- tethered/remote worker 473
- text database 831
- theory of
 - portable PKI 397
 - reasoned action (TRA) 677
- therapeutic
 - guidelines 99
 - vaccines 114
- third generation 551

Index

- third-generation (3G)
 - networks 709
 - wireless technologies 148
 - partnership project (3GPP) 173, 180
- three-tier architecture 249
- time
 - aligned idle period downlink (TA-IPDL) 44
 - consciousness 661
 - dependent CVP 339
 - estimation policies 339
 - difference of arrival (TDOA) 3
 - division
 - duplex (TDD) 174
 - multiple access (TDMA) 144, 145
 - independent CVP 341
 - estimation policies 341
- TKIP (see temporal key integrity protocol)
- TLS (see transport layer security)
- total
 - access communications system (TACS) 144
 - cost of ownership (TCO) 769
- touch screens 411
- toxicokinetics 114
- TPL (see Trust Policy Language)
- TRA (theory of reasoned action)
- track-and-trace technologies 381, 382
- tracking user's preferences 538
- traditional summarization 250
- traffic information system 608
- training 609
- transactions 695
- transition plan 593
- transport layer security (TLS) 397
- travel recommender systems (TRSs) 812
- travelling 593
- TRSs (see travel recommender systems)
- Trust Policy Language (TPL) 614
- TU (see telematic unit)
- types of LBSs 22

U

- ubiquity 633, 697
- UIR (see updated invalidation report)
- ULDs (see unit load devices)
- ultra wide band 446
- UML (see Unified Modeling Language)
- UMTS (see universal mobile telecommunication system)
- unconnected mobile tourist 814
- unicasting (one-to-one) technology 755

- unification of approaches 280
- Unified Modeling Language (UML) 786
- uniformity 662
- unit load devices (ULDs) 744
- universal description, discovery, and integration
 - 65, 242, 547
- universal mobile telecommunication system (UMTS) 149, 421
- unpredictable object movement modeling 281
- unsupervised agent transport 372
- update reports (URs) 133
- updated invalidation report (UIR) 133
- URs (see update reports)
- usability 645, 659
- user
 - experience 648
 - feedback 535
 - identification issues 230
 - login 616
 - profiles 598
 - profiling 529
 - requirements 600
 - preferences 539
- utilities management systems 794

V

- value
 - creation in Mobile Internet 423
 - propositions 425
 - added mobile services 633
- variations in mobility 319
- VC (see virtual clock)
- vertical market composition agent 243
- VGA high-resolution screen 98
- videoconferencing 149
- viral marketing 647
- virtual
 - clock (VC) 311
 - private networks (VPNs) 421, 474
- visiting 593
- Voiceover
 - IP (VoIP) 149, 436
 - Wi-Fi (VoWiFi) 436
 - Wireless
 - IP (VoWIP) 436
 - LAN (VoWLAN) 436
- voice quality 144
- VoIP (see Voiceover IP)
- VoWiFi (see Voiceover Wi-Fi)
- VoWIP (see Voiceover Wireless IP)

VoWLAN (see Voiceover Wireless LAN)
VPNs (see virtual private networks)

W

W3C (see World Wide Web Consortium)
WAN (see wide area network)
wandering 593
WAP (see wireless application protocol)
 i-mode 405
 Gateway 86
 banking 424
WAPs (see wireless access points)
W-CDMA (see wideband code division multiple access)
Web
 blog 668
 browsing 248
 Ontology Language (OWL) 55
 servers 408
 service (WS) 56, 88, 481, 544, 570
 ontology 63
 Services Description Language (WSDL) 547
 services-based operator 107
 site database 831
WEP (see wired equivalent privacy)
white pages 65
wholesale and retail CRM applications 774
Wi-Fi 99, 436, 445
 wireless fidelity 695
 protected access (WPA) 396, 732
 telephony 436
wide area network (WAN) 406, 696
wideband code division multiple access (W-CDMA) 158
WiMAX 446
wired
 equivalent privacy (WEP) 395
 networks 407
wireless 35
 access
 points (WAPs) 395
 protocol (WAP) 709
 and wired networks 414
 application protocol (WAP) 146, 247, 405, 424, 464, 488, 494, 551, 584, 585
 authentication and privacy infrastructure 488
 carriers 467
 communication 132, 143
 computing 466
 connectivity 735

data privacy enhancements 732
handheld 553
infrastructure 735
local area network (WLAN) 184, 322, 394, 406, 489, 549, 724, 730, 755
 architecture 729
Markup Language (WML) 247, 484
market 722
metropolitan area networks 724
miniature devices for CHF patients 120
network interface card (WNIC) 490
networking evolution 551
networks 184, 406
personal area network (WPAN) 323, 406
policy 771
private area networks 394
technologies 144
Voiceover IP (WVoIP) 436, 735
wide area network (WWAN) 394, 406, 724
USB 446
WLAN (see wireless LAN)
WML (see Wireless Markup Language)
WNIC (see wireless network interface card)
work settings 70
workforce management 36
World Wide Web Consortium (W3C) 55
WPA (see Wi-Fi protected access)
WPAN (see wireless personal area network)
wrist BP monitors 119
WS (Web service)
WSDL (see Web Services Description Language)
WVoIP (see wireless Voiceover IP)
WWAN (see wireless wide area network)

X

XML (see extensible markup language)
XSLT (see extensible style sheet transformation)

Y

Yellow Pages 65
yield 40

Z

ZigBee 445