Undergraduate Lecture Notes in Physics

Christo Papadopoulos

Solid-State Electronic Devices

An Introduction

Springer

Undergraduate Lecture Notes in Physics

For further volumes:
http://www.springer.com/series/8917

Undergraduate Lecture Notes in Physics (ULNP) publishes authoritative texts covering topics throughout pure and applied physics. Each title in the series is suitable as a basis for undergraduate instruction, typically containing practice problems, worked examples, chapter summaries, and suggestions for further reading.

ULNP titles must provide at least one of the following:

- An exceptionally clear and concise treatment of a standard undergraduate subject.
- A solid undergraduate-level introduction to a graduate, advanced, or non-standard subject.
- A novel perspective or an unusual approach to teaching a subject.

ULNP especially encourages new, original, and idiosyncratic approaches to physics teaching at the undergraduate level.

The purpose of ULNP is to provide intriguing, absorbing books that will continue to be the reader's preferred reference throughout their academic career.

Christo Papadopoulos

Solid-State Electronic Devices

An Introduction

Springer

Christo Papadopoulos
Electrical and Computer Engineering
University of Victoria
Victoria, BC, Canada

ISSN 2192-4791 ISSN 2192-4805 (electronic)
ISBN 978-1-4614-8835-4 ISBN 978-1-4614-8836-1 (eBook)
DOI 10.1007/978-1-4614-8836-1
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2013949154

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

To George, Meropi, and those who have given more than they have taken.

Preface

This book provides an elementary treatment of the solid-state electronic devices that are the foundation of modern electronic systems and information technology. The concepts and devices presented underlie semiconductor integrated circuit technology and virtually all of the industrial and consumer electronics products that are now so ubiquitous and indispensable to everyday life. From an academic point of view, electronics is a subject that today impacts and is important to almost every field of science and engineering due to its broad and cross-cutting interdisciplinary nature.

The aim of this textbook is to provide a concise exposition on the main electronic devices upon which most modern electronics is based, in addition to emerging trends and applications of this technology. The main audience is undergraduate and beginning graduate students in engineering and applied physics. However, the book was written so that anyone with first/second year university level physics and mathematics should be able to understand the material. Physics, chemistry, biology, engineering, and related fields are becoming more and more intertwined, and the text tries to use an approach that maintains accessibility to a wide audience. Readers of this book can expect to derive a solid foundation for understanding today's electronic devices and also be prepared for future developments and advancements in this far-reaching area of science and technology.

The contents are largely based on an undergraduate course I have taught at the University of Victoria and developed further during a brief study leave. I have attempted to constrain the treatment to those structures/devices that are most relevant to modern electronics. As a result, several topics and devices have not been included so as not to detract from the fundamental concepts. In several instances, straightforward extensions of the ideas presented can be found in one of the many excellent more advanced and comprehensive references available.

Chapters 1–4 cover the basic solid-state electronic structures and devices required for a one-term course. Chapter 5 on emerging areas and devices provides a glimpse into the future of electronics and also shows how the core electronics material of the previous chapters is being extended into new directions and leading to novel applications in many diverse areas. Appendix A provides a review of the

relevant physics of semiconductors needed to understand electronic devices and can be used as a starting point or simply referred to when necessary.

I would like to thank Christopher T. Coughlin and Ho Ying Fan from Springer for their support and patience during the writing process. I also thank Spyrithoula-Alida Gountas for assisting with several of the figure illustrations. As always, the support of my family has been a key component that inspired this book and helped it come to fruition. Lastly, I would like to acknowledge Jimmy Xu who first introduced me to the rich and exciting field of electronic devices while I was an undergraduate student at the University of Toronto.

Victoria, BC, Canada Christo Papadopoulos

"The Lioness and the Fox meeting together, fell into discourse; and the conversation turning upon the breeding, and the fruitfulness of some living creatures above others, the Fox could not forbear taking the opportunity of observing to the Lioness, that for her part, she thought Foxes were as happy in that respect as almost any other creatures; for that they bred constantly once a year, if not oftener, and always had a good litter of cubs at every birth: and yet, says she, there are those who are never delivered of more than one at a time, and that perhaps not above once or twice through their whole life, who hold up their noses, and value themselves so much upon it, that they think all other creatures beneath them, and scarce worthy to be spoken to.

The Lioness, who all the while perceived at whom this reflection pointed, was fired with resentment, and with a good deal of vehemence replied: what you have observed may be true, and that not without reason. You produce a great many at a litter, and often, but what are they? Foxes. I indeed have but one at a time, but you should remember, that this one is a Lion."

Aesop[1]

[1] *Fables of Aesop and others*, translated by Samuel Croxall.

Contents

Chapter 1
Introduction

"[The] great accomplishments of experimental science were achieved by men of many types: patient, persistent, intuitive, energetic, lazy, lucky... Most had in common only a few things: They were honest and actually made the observations they recorded...in a form permitting others to duplicate the experiment or observation."

C. Kittel, *Mechanics*

Electrical systems are ubiquitous in modern technology and have a major socioeconomic influence on the world. It is difficult to find an aspect of human life that they have not impacted: Electrical systems are the foundation of energy distribution networks, electric machines, lighting, sensors, control systems, manufacturing and, largely through the development of the integrated circuit, are fundamental to applications including computers, consumer electronics, telecommunications (the Internet, mobile devices, etc.), and information processing technology in general. These advances have been enabled by a detailed understanding of the electronic properties of materials.

1.1 Electronics

1.1.1 Brief History of Electronic Materials

The study of the electrical properties of materials is an ancient endeavor.[1] However, the systematic knowledge of how electrons behave inside materials and particularly the widespread use of electrical phenomena had to await much more recent advances

[1] Thales of Miletus is said to have commented on the ability of amber to attract other objects ca. 600 BC. Some time later, ca. 1600 AD, William Gilbert used the Greek word for amber,

C. Papadopoulos, *Solid-State Electronic Devices: An Introduction*,
Undergraduate Lecture Notes in Physics, DOI 10.1007/978-1-4614-8836-1_1,

in the basic understanding of electromagnetic fields and particles. It is interesting to perhaps contrast this with other types of materials and processes that were developed and applied to a great extent well before a modern scientific understanding arose-fire, concrete, iron/steel, glass, wood/textiles, hydraulics, fuel, etc. Historically, the chemical, mechanical, gravitational, optical, and thermal potentials of physical systems were much easier to control and manipulate than "catching lightning in a bottle" in order to use electrical forces. Thus, innovations such as aqueducts, great buildings, bridges, mechanical clocks, glass windows and vessels, mirrors and lenses, metallurgy, ships, and wine[2] all have long histories. It is not surprising therefore that when conditions became ripe for the industrial revolution in the 1700s, it was driven by steam, coal and gas, chemical production, mining, and several other industries whose foundation had been laid over thousands of years—electricity was notably and, as we have seen, necessarily, absent.

The first major strides towards a rational understanding of electricity and magnetism began to appear around the latter half of the eighteenth century and onwards with the pioneering experiments and deductions of Cavendish, Coulomb, Gauss, Oersted, Ampère, Faraday, and many others. These important results would eventually lead Maxwell to a complete classical electromagnetic field theory by 1864. Another very important practical element that enabled early progress into the application of electricity was the availability of a stable electrical power source in the form of the electrochemical cell or battery, first demonstrated by Volta in 1800. Once again it is not surprising that a chemical approach to generating an electrical potential was the first to be developed, making use of prior knowledge gained from chemistry and chemical reactions. Commercial electricity generation in the form of central plants would only become commonplace almost 100 years later. The battery allowed electric currents to flow in a continuous manner through a circuit and electronic transport through various materials could be studied. The thermoelectric effect discovered by Seebeck in 1821 provided another stable power source. By 1826 these advances had allowed Ohm to perform experiments which led to his now famous law relating current and voltage through a resistor and in 1841 Joule had established the amount of heat dissipated during such current flow.

By the close of the nineteenth century the state of science and engineering was beginning to enter what we could call the modern era: quantum theory would soon lead to a quantitative understanding of the atomic nature of matter and elementary particles, (special) relativity would put electricity and magnetism on a firm theoretical foundation, and many important experimental and theoretical results appeared including the identification of the electron by Thomson around 1897 and the precise measurement of its quantized charge by Millikan in 1909. This led to the emergence of solid-state physics and a much deeper understanding of the electronic properties of solids, including why some (e.g., metals) conduct electricity so well while others (e.g., glass) are essentially insulating.

ēlektron, to describe the field surrounding a charged object, from which we derive the modern terms electron, electric field, electronics, etc.

[2] Falernian wine, for example, was prized in ancient Rome for its high-quality and aging ability, with superb vintages lasting several decades or more.

This tremendous foundation, established within the last 200 years, forms the pillars of modern electronic devices. It is important to emphasize once more that the work of early investigators into the science of electricity and magnetism and materials was by no means trivial: As alluded to above, being able to control the distribution of charges in order to perform a useful task is a very difficult problem in practice as electricity is in general a fleeting phenomenon, compared to, say, gravity or the potential energy in a spring or chemical reaction. Materials are neutral under most conditions one would normally encounter, and ascertaining the nonequilibrium electrical behavior of a system required the superb skill and groundbreaking insight of many individuals.

This textbook is essentially about the fruits of these labors, which have led to what we refer to as modern electronics, the building blocks of modern society and information technology. Today, we can say in many ways that pen and paper have largely been replaced by the electron. More generally, the workhorse of almost every endeavor is now based on using electricity to power and control devices that convert and process information in one form or another. Alongside the "ancient" applications of structures, mechanics, chemistry, and optics, electronics has in the last 100 years transformed society more than any other technology.

1.2 The Importance of Semiconductors

The earliest electronic devices (beyond the simple conducting wire or resistor) employed mechanical and/or chemical elements in their design, combining mechanical or active chemical components with electronic ones, and were mainly concerned with making viable telecommunications possible. For example, in the early 1800s electrochemical telegraphs demonstrated that they could transmit information over several kilometers by passing electric current through acidic baths at the receiver's end in order to evolve gases and hence create a visual signal. Electromechanical telegraphs were developed shortly thereafter and consisted of indicators that deflected in response to electrical signals (due to a charged wire, for example).

An important extension of these ideas was enabled by concurrent progress on understanding the nature of electricity and magnetism: By employing electromagnetic induction and in particular the recently invented electromagnet (or coil), magnetic fields could be sensed and generated for sending and receiving of information. Around the middle of the nineteenth century due to the work of Morse and others, such electromagnetic telegraphs had progressed to the point that practical commercial systems had been established connecting cities across continents, and what is often termed the era of *electrical telegraphy* was well under way.[3]

[3] An interesting take on the early electric telegraph network and its applications can be found in T. Standage, *The Victorian Internet*, 1998, where telegraphy is shown to have much in common with today's Internet.

One problem that early electrical telecommunications presented was the signal loss that occurred over long-distance lines. The electromagnetic relay was an electrically actuated mechanical switching device developed in 1835 by Henry that would improve long-distance communications by allowing amplification of an incoming signal between sender and receiver via connection to an intermediate power source (battery), thus acting as a repeater. Such relays would later become important for switching in telephone networks. These functions, viz., (signal) *amplification* and *switching*, define two of the fundamental applications of electronic devices.

Examples of other early electrical devices that would have a major impact on society include the electric motor that began being developed during Faraday's pioneering work on electromagnetism in the 1820s[4] and would be improved upon and reach its more familiar form by the end of the century. The electric motor transformed the operation of most industrial processes and an immense range of other applications and household products requiring work via mechanical motion—it has proven to be one of the most useful applications of electricity. Another important application that has become virtually indispensable to modern life is the electric light source: The modern lighting industry emerged with the incandescent light bulb that was commercialized based on the designs of Swan and Edison starting around 1880.[5] The development of efficient artificial electric light sources continues to this day.

As the use of electric telegraphy and electricity matured, other types of telecommunication technology began to be developed. In particular, telephony (essentially modifying an electric telegraph to transmit audio information via electromagnetically induced currents) that was kick-started by Bell in 1876[6] and radio (or wireless telegraphy) developed by Marconi and others in the 1890s would eventually replace the telegraph as the most widely used modes of long-distance communication by the middle of the twentieth century. These new technologies also brought with them additional challenges and thus new devices were developed as a result.

In 1877 the carbon microphone (a.k.a., the carbon transmitter) was first demonstrated by Edison and subsequently improved by Hughes in 1878. In this device, the pressure from sound waves would cause the resistance through a collection of small carbon particles or granules to change and this variable resistance would become widely used for many decades to convert sound into the electrical signal transmitted by the telephone and was the precursor to later iterations for recording sound in modern audio equipment. In addition, carbon microphones were also used as amplifiers in early telephone network repeaters in order to boost signals over

[4] Electric generators, or the reverse of the motor, which would become the basis of the electrical power grid, began being developed around same time with the designs of Tesla in particular making important strides towards modern power generation and transmission by the end of the nineteenth century.

[5] Carbon-arc (gas-based) lighting was also developed commercially around the same time by the Thomson-Houston Electric Company, which would subsequently merge with Edison General Electric to form General Electric in 1892.

[6] Bell's original telephone patent was entitled "Improvement in Telegraphy"; US patent 174,465.

long-distance lines by converting the incoming electrical signal to a mechanical vibration (via an electromagnetic receiver) and coupling it back into a microphone to generate a stronger signal.

The detection of electromagnetic (EM) radiation was a key element in the development of radio technology. Following the pioneering observation of EM waves by Hertz in 1888, the so-called coherer was the standard detector used in early radio systems. This device consisted of a glass tube that was filled with small metallic particles (iron filings) whose resistance would change upon exposure to EM radiation (thought to occur via a melding or coherence of the particles). The coherer was soon to be replaced by a more reliable device that would begin an important era in the history of electronics, i.e., the *vacuum tube*. By modifying Edison's incandescent light bulb design, Fleming developed the first diode or two-terminal vacuum tube device that could both detect radio waves and rectify signals, which he patented in 1904. This was soon to be followed by a three-terminal vacuum tube device or triode, which built upon de Forest's Audion tube of 1906. These devices were the first electronic amplifiers and heralded a revolution in electronics by allowing the detection and processing of (weak) signals in ways not previously possible. Some of the major advances enabled included transcontinental and transoceanic telephony (by allowing reliable repeaters), radiotelephony (as opposed to simple telegraphy), radio and television broadcasting, and the building blocks for a multitude of analog electronic and early computational devices.

By the first half of the twentieth century, the transformation of electronics from its early hybrid mechanical/chemical forms to all-electrical devices was complete.

1.2.1 Solid-State Electronics

All of the progress and success enabled by electronics by the turn of the twentieth century led to a continual search and push for devices with better and better performance (a trend that has continued up to the present day). Notwithstanding these achievements, it is perhaps fair to say that up to this point the devices, from an electronic materials viewpoint, were quite simple, whether one considers electromechanical devices essentially consisting of coils and wires or vacuum tubes consisting of metallic electrodes in a light bulb-type glass housing.

Despite their sometimes complex inner workings these devices consisted of just two electronic materials: metallic conductors (e.g., copper) and insulators (e.g., glass). A third type of material, known as a *semiconductor*,[7] would come to dominate electronics in the latter half of the twentieth century and defines modern electronic devices—the topic of this book.

Semiconductor crystals, at first glance, are not much different than any other periodic arrangement of atoms. However, with advances in the microscopic

[7] A semiconductor can be thought of as a special type of insulator, as discussed in Appendix A.

understanding of matter due to quantum mechanics in the 1920s and 1930s, it was shown that their electronic properties were highly tunable, with behavior ranging from almost completely insulating to almost metallic. With this new degree of freedom, along with advances in material processing, one could begin to create complex *solid-state* devices that did not contain any moving parts or vacuum tubes. These devices could thus be faster, smaller, more reliable, and consume less energy while still maintaining the complex functions of earlier mechanical or vacuum devices.

Once semiconductor physics and processing had advanced, the middle of the twentieth century brought about groundbreaking advances in the form of solid-state analogues to the diode and triode. In particular, the landmark demonstration of the transistor at Bell labs in 1947 is widely considered the beginning of the modern era in electronics and information technology.

1.2.2 The Integrated Circuit and Information Technology

Following the establishment of discrete solid-state devices it became clear to some that finding an effective way to connect or *integrate* these devices together into large and complex circuits would be a major factor in the continued progress of electronics. This was made plainly obvious by considering that state-of-the-art computers around 1950 such as the ENIAC, which contained tens of thousands of discrete devices, had to be stored in very large rooms.

In the late 1950s two pioneers, Kilby (Texas Instruments) and Noyce (Fairchild Semiconductor and later cofounder of Intel[8]), determined the fundamental ideas and processes that would be used to integrate many electronic devices and large circuits onto a single piece of semiconductor crystal. The resulting *integrated circuit* or IC, based primarily (up to now) on the semiconductor *silicon*, is perhaps the most important technological process invented in modern history. Silicon has proven itself to be a truly superb material that is also very versatile (and very resilient); it really defines the modern age of information. Today billions of electronic devices are routinely integrated together in small chips to form powerful circuits for applications ranging from mobile phones to industrial process control and automatic piloting of airplanes. This technology has enabled the Internet and the vast amount of digital information now being sent over modern fiber-optic, wired, and wireless networks can be processed efficiently due to the ICs that power the devices generating and receiving this information. Electronics and telecommunications have thus continued their symbiotic relationship that began with the telegraph.

The immense applications of electronics have led to it becoming the world's largest industry with global sales reaching approximately 2 trillion USD in 2012.

[8] The original Intel business plan is remarkable for its efficiency and was apparently prepared by Noyce in less than 1 h. It is displayed in its original form at the end of this chapter.

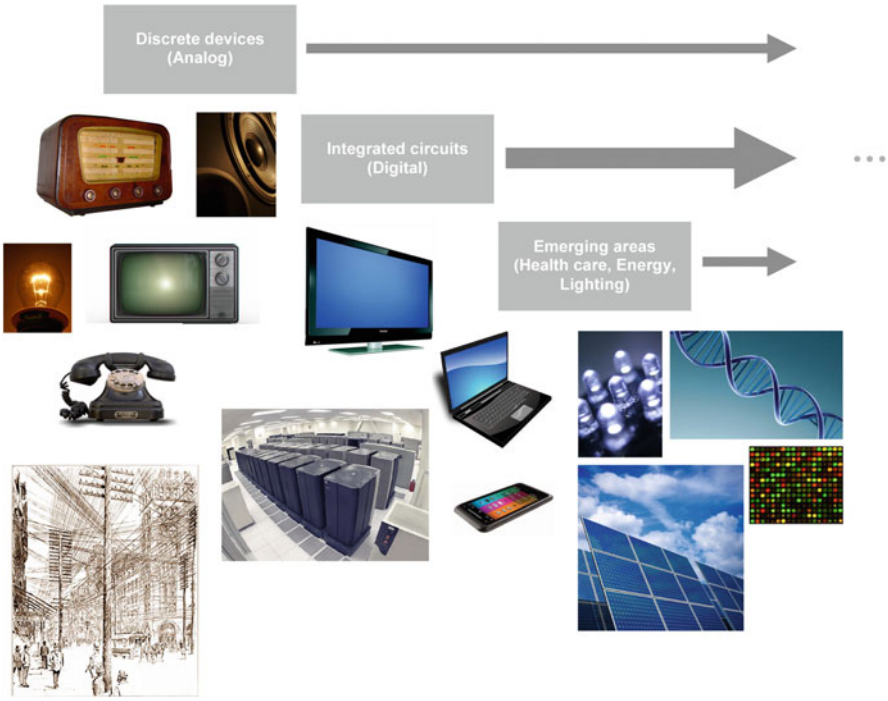


Fig. 1.1 Evolution of solid-state electronics and emerging growth areas

1.2.3 Emerging and Growth Areas

While the electronics industry continues to grow at a strong pace, new areas of application and opportunities are also emerging. The evolution of solid-state electronics thus far with a view towards the future is shown schematically in Fig. 1.1. Digital applications and ICs should continue to drive growth for some time. In addition to the established areas, the current push towards renewable energy sources and energy-efficient lighting, along with biomedical applications, appear to be active growth opportunities for electronics moving forward. Others are sure to emerge in this very rich field.

1.3 Overview of Text and Suggestions for Use

Chapters 2, 3, and 4 form the core material of modern solid-state electronic devices and their most important applications. Chapter 5 extends the standard semiconductor device knowledge into emerging areas of electronics and other fields. Although the final chapter is not strictly required to understand most electronic devices, it is

definitely the most important for understanding some of the future directions the semiconductor industry is taking, in addition to showing that this field is now very interdisciplinary and extends well beyond "standard" electrical engineering and electronics.

Appendix A is meant to provide the minimum background physics knowledge required in order to effectively use this book. Readers who are somewhat unfamiliar with quantum mechanics and/or semiconductors should first complete the short primer in the appendix before moving on to Chap. 2. Alternatively, those well versed in these concepts may simply refer to the appendix as required.

References

1. Elliot, R.S.: Electromagnetics: History, Theory, and Applications. IEEE Press, New York (1993)
2. Huurdeman, A.A.: The Worldwide History of Telecommunications. Wiley, Hoboken (2003)
3. Ng, K.K.: Complete Guide to Semiconductor Devices, 2nd edn. Wiley Interscience, New York (2002)

Intel Business Plan, July 1968

Source: Intel Museum.

The following plan was used to raise $2.5 million in venture capital in less than 2 days.

The company will engage in research, development, adn manufacture and sales of integrated electronic structures to fulfill the needs of electronic systems manufacturers. This will include thin films, thick films, semiconductor devices, and other solid state componenat used ih hybrid and monolithic integrated structures.

A variety of processes will be established, both at a laboratory and production level. These include crystal growth, slicing, lapping, polishing, solid state diffusion, photolithographic masking and etching, vacuum evaporation, film deposition, assembly, packaging, and testing, as well as the development and manufacture of special processing and testing equipmentrequired to carry out these processes

Products may include dioded. transistors, field effect devices, photo sensitive devices, photo emitting devices, integrated circuits, and subsysteme commonly referred to by the phrase "lagge scale integration" Principql customers for these products are expected to be the manufacturers of advanced electronic systems for communications, radar, control and data processing. It is anticipated that many of these customers will be located outside California.

Chapter 2
Junctions and Diodes

". . . the totality is not, as it were, a mere heap, but the whole is something besides the parts . . ."

Aristotle, *Metaphysics*

A junction between two dissimilar materials is shown schematically in Fig. 2.1 below. Such a junction can take many forms; however, its defining characteristic is the asymmetry that exists upon crossing the junction from one side to the other. This is the essence of the original term used to describe such two-terminal junction devices, viz., *diode*, literally meaning two different paths or roads. Historically, the earliest solid-state junctions were formed in the latter half of the nineteenth century by mechanically pressing a sharp metallic object (e.g., wire) against a semiconducting material; the so-called cat's whisker or point-contact diode.[1] Similar techniques were also used to create junctions between two different semiconductor crystals and between thin layers of metals and semiconductors. Vacuum tube diodes on the other hand, developed around the same time, obtain their asymmetry via the heating of one electrode (the cathode) relative to the other (the anode).[2] The cathode material is chosen such that it readily emits electrons inside the vacuum tube when heated (thermionic emission).

The asymmetry of a junction leads to currents that depend on the polarity of the bias applied to the junction; in other words they possess an asymmetric *I–V* characteristic. Such junctions can be engineered to create a wide range of solid-state electronic devices. It is also often necessary, and very important in practice, to mitigate the effects of junctions when making electrical contacts or interconnecting

[1] Often referred to simply as a crystal diode (detector).

[2] Mercury-arc valves were another early type of diode based on a liquid mercury cathode and carbon anode. In this case, electrons are preferentially emitted from the cathode upon formation of a flame or arc discharge (through the contained mercury vapor).

C. Papadopoulos, *Solid-State Electronic Devices: An Introduction*,
Undergraduate Lecture Notes in Physics, DOI 10.1007/978-1-4614-8836-1_2,

Fig. 2.1 General schematic of two different materials in intimate contact. A junction is formed at the interface between A and B

multiple devices. In this chapter, we consider the main types of solid-state junctions and their applications based on diodes. Such junctions form the basic building blocks of all solid-state electronic devices and thus their properties will play a critical role throughout the subsequent chapters.

2.1 *pn* Junctions

The first type of junction we consider is between two regions in a semiconductor having different doping type. The earliest experiments on *pn* junctions are attributed to Ohl working at Bell labs in 1940 on the properties of silicon crystals.[3] Following this discovery, the theoretical understanding of such junctions was largely developed by Shockley.[4]

2.1.1 Thermal Equilibrium and the Built-In Potential

We use the fact that the Fermi level must be constant in thermal equilibrium to construct the band edge diagram for junctions between different materials. In the case of a *pn* junction, before the two materials are in contact we have the situation depicted in Fig. 2.2a: E_0 is the vacuum level and represents the energy of free electrons (in other words, carriers that are no longer bound to the material). The difference between the Fermi level and E_0 is called the *work function*, $q\Phi$, of the material. We also define the difference between the conduction band edge, E_c, and the vacuum level as the *electron affinity*, qX, which is constant for a given semiconductor.

If the two regions now come into contact and combine to form a junction, the large carrier concentration gradients that exist at the interface will cause a transfer of carriers in order to align the Fermi levels and achieve thermal equilibrium. This leads to a *depletion layer* or *space-charge* region near the interface, caused by uncompensated (fixed) impurity ions.[5] The thermal equilibrium band edge

[3] The discovery of an unintentional junction in a rod of crystalline silicon by Ohl led to the terms "n-type" and "p-type" being coined to describe the different doping on either side of the junction. Ohl's experiments also resulted in the demonstration of the first *pn* junction solar cell.

[4] W. Shockley, Bell Syst. Tech. J. **28**, 435 (1949). This was a seminal paper in the history of semiconductor electronics and forms much of the foundation on which Chapters 2 and 3 are based.

[5] To equalize the Fermi levels there will be a net transfer of electrons toward the p-type region and recombination with holes results in the space-charge layer at equilibrium.

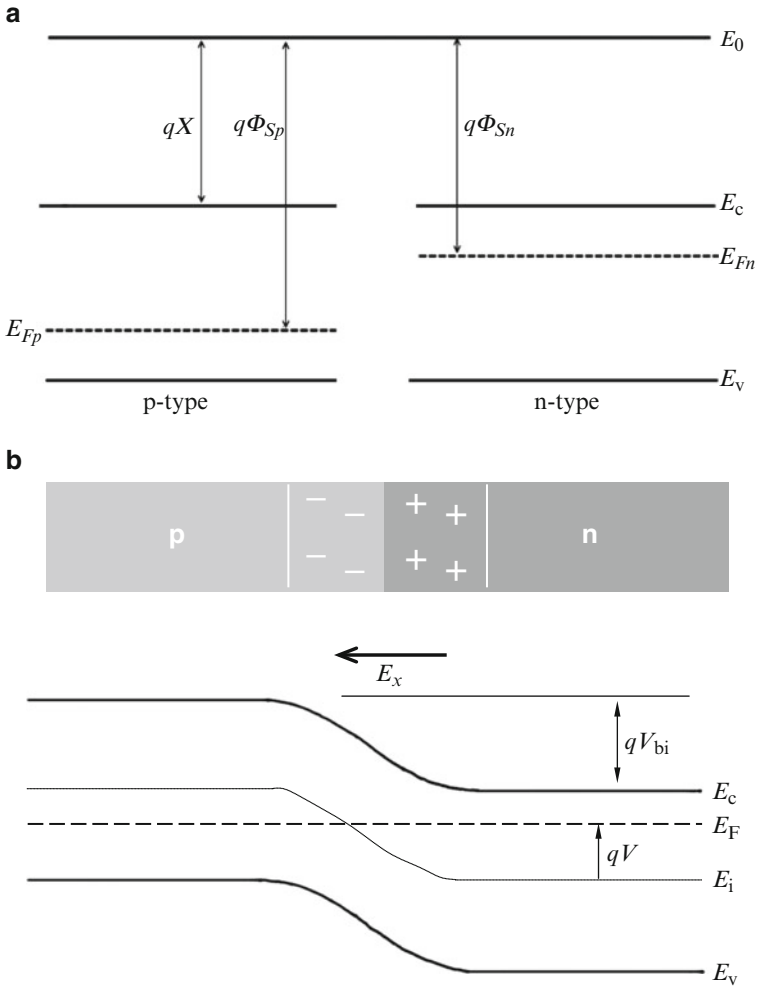


Fig. 2.2 (**a**) Band edge diagrams for isolated p-type and n-type semiconductors, including the vacuum reference energy level, E_0. (**b**) Thermal equilibrium *pn* junction band edge diagram, characterized by the built-in potential barrier V_{bi}. The potential at any point in the junction can also be defined by V as shown. (The electric field in the x-direction (E_x) always points "uphill" with respect to the conduction band edge.) A schematic of the built-in charge of the depletion layer appears above the band edge diagram. The interface should be visualized (in both diagrams) as a two-dimensional sheet or plane (coming out of the page) where the three-dimensional crystals meet. Note: To qualitatively sketch junction band edge diagrams one should start by drawing the thermal equilibrium Fermi level as a horizontal line, followed by sketching the band edge diagrams away from the interface on either side (i.e., in the bulk regions, where the bands are flat). Lastly, the conduction and valance band edges should be smoothly joined together in the junction region where the two materials meet

diagram for the junction thus has the form depicted in Fig. 2.2b. The resulting electric field is characterized by the *built-in potential*, V_{bi}, at equilibrium. In thermal equilibrium the built-in field exactly balances the tendency of carriers to diffuse across the junction so that there is no net current flow.

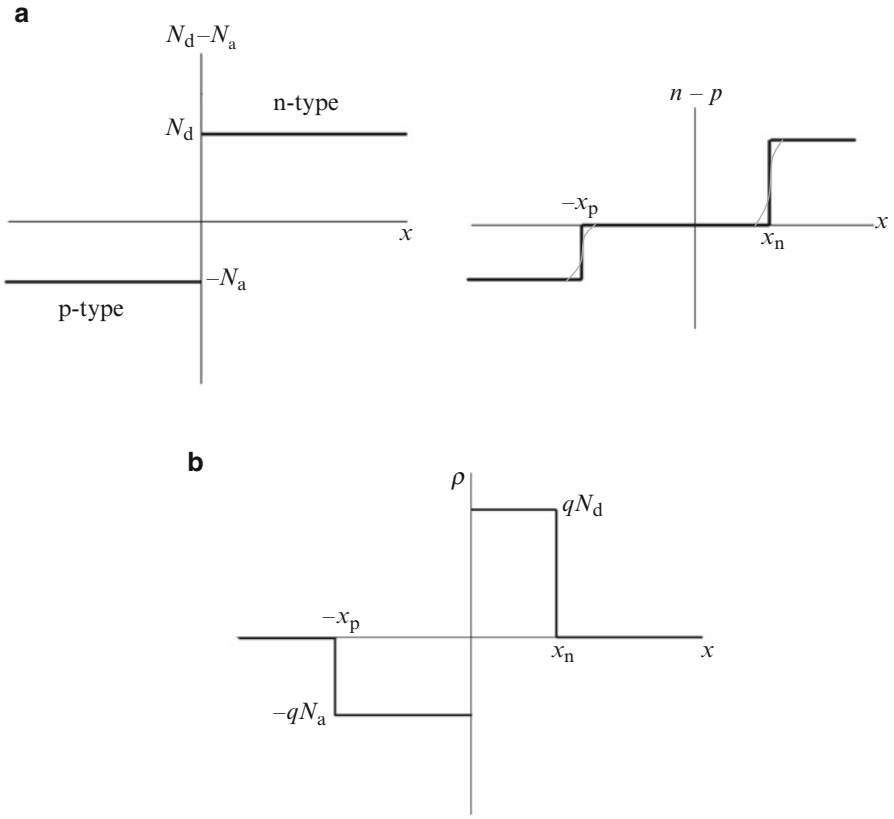


Fig. 2.3 (**a**) Depletion approximation for an abrupt *pn* junction: Doping levels for the step junction (*left*) and resulting mobile carrier concentrations within the depletion approximation (*right*; *thin line* shows behavior of exact solution). (**b**) Charge density for the step *pn* junction under the depletion approximation. (**c**) Electric field across step *pn* junction for charge distribution of Fig. 2.3b. (The field is negative because it is pointing the negative x-direction.) (**d**) Electric potential variation across *pn* junction for field shown in Fig. 2.3c (Note that the reference point for the potentials can be shifted without loss of generality)

To solve for the potential barrier V_{bi} one usually employs the *depletion approximation*, which ignores any contribution to the space charge from free electrons and holes, as depicted in Fig. 2.3a. This is a very good approximation due to the exponential dependence of free carrier concentration with Fermi-level position; in the vicinity of the junction the magnitude of $(E_F - E_i)$ becomes small (Fig. 2.2b) and hence the free carrier concentration rapidly decreases. Note that we are also assuming (as in Fig. 2.2) what is known as an abrupt or step junction, in that the doping levels in the p- and n-type materials are constant away from the junction and abruptly change at the junction interfacial region.[6]

[6] The interface or plane where the doping type changes is known as the *metallurgical junction*. The same terminology is used to describe the physical or material transition interface in any type of solid-state junction.

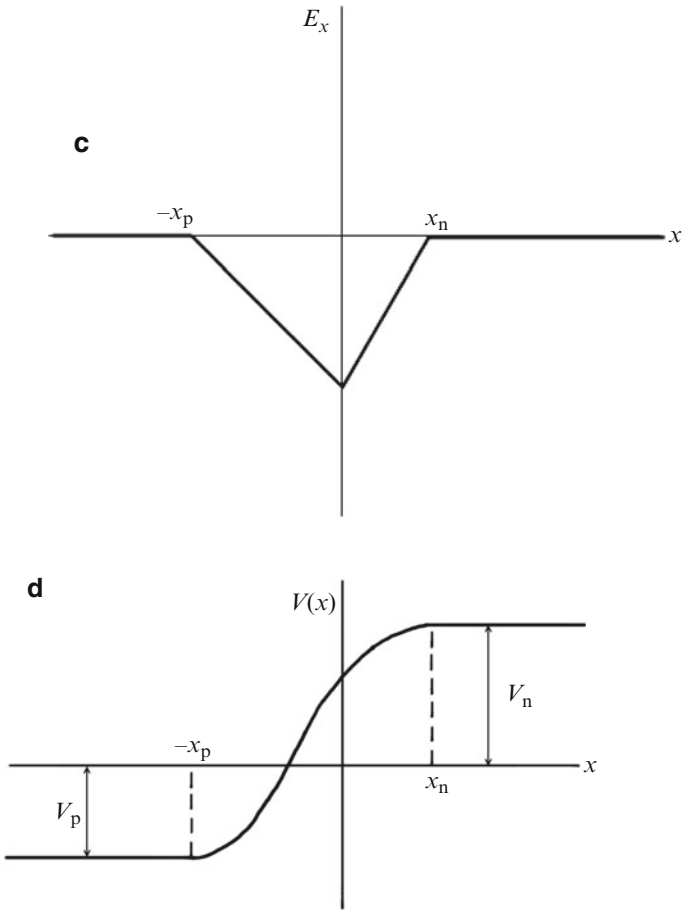


Fig. 2.3 (continued)

This allows us to write Poisson's equation in the space-charge region as[7]

$$\frac{d^2V}{dx^2} = \frac{-q}{\varepsilon_s}(N_d - N_a) \tag{2.1}$$

where the charge density has the form shown in Fig. 2.3b. Thus in the n-type material ($x > 0$) Poisson's equation becomes

[7] Throughout this book the device analysis and descriptions are mainly constrained to one spatial dimension (e.g., x) in order to emphasize and develop an understanding of the basic principles of solid-state electronic devices. The 1D picture will be extended to include additional dimensions/effects when necessary.

$$\frac{d^2V}{dx^2} = -\frac{dE_x}{dx} = \frac{-qN_\mathrm{d}}{\varepsilon_\mathrm{s}} \tag{2.2}$$

which can be integrated to the edge of the depletion region at x_n, where the material becomes neutral and the field vanishes:

$$E_x = -\frac{qN_\mathrm{d}}{\varepsilon_\mathrm{s}}(x_\mathrm{n} - x), \quad 0 \le x \le x_\mathrm{n} \tag{2.3}$$

Similarly, in the p-type region we can find

$$E_x = -\frac{qN_\mathrm{a}}{\varepsilon_s}\left(x + x_\mathrm{p}\right), \quad -x_\mathrm{p} \le x \le 0 \tag{2.4}$$

Thus the field is negative throughout the depletion region and varies linearly with x, reaching a maximum at $x = 0$ (the metallurgical junction) as shown in Fig. 2.3c.

The electric field must also be continuous at the interface, $x = 0$, so that

$$N_\mathrm{a}x_\mathrm{p} = N_\mathrm{d}x_\mathrm{n} \tag{2.5}$$

This equation tells us that the width of the depletion region on either side of the junction varies inversely with doping concentration. In other words, the depletion region extends primarily into the side which has the *lightest* doping level. Note that Eq. (2.5) is also another statement of (global) space-charge neutrality (see Appendix A, Eq. (A.31)).

Since

$$E_x = -\frac{dV(x)}{dx}$$

we can integrate the above expressions for the field to obtain the potential variation across the junction. This gives

$$\begin{aligned} V(x) &= V_\mathrm{n} - \frac{qN_\mathrm{d}}{2\varepsilon_\mathrm{s}}(x_\mathrm{n} - x)^2, \quad 0 \le x \le x_\mathrm{n} \\ V(x) &= V_\mathrm{p} + \frac{qN_\mathrm{a}}{2\varepsilon_\mathrm{s}}\left(x + x_\mathrm{p}\right)^2, \quad -x_\mathrm{p} \le x \le 0 \end{aligned} \tag{2.6}$$

where V_n and V_p are the potentials at the neutral edges of the depletion region on either side of the junction as shown in Fig. 2.3d. The parabolic dependence of the potential is to be expected upon integrating Poisson's equation twice for a constant charge density and also describes the curvature of the band edges across the space-charge region.[8]

[8] Electron energies (and hence the energy band edges) are given by $E = -qV$. See Fig. A.13a in Appendix A for the general relation between the band edges and electric field.

The potentials at the boundaries of the junction can be found by noting

$$n = n_\mathrm{i}\ \exp\left(\frac{qV_\mathrm{n}}{k_\mathrm{B}T}\right);\ \ p = n_\mathrm{i}\ \exp\left(\frac{-qV_\mathrm{p}}{k_\mathrm{B}T}\right) \tag{2.7a}$$[9]

which gives

$$\begin{aligned} V_\mathrm{n} &= \frac{k_\mathrm{B}T}{q}\ln\frac{N_\mathrm{d}}{n_\mathrm{i}} \\ V_\mathrm{p} &= \frac{-k_\mathrm{B}T}{q}\ln\frac{N_\mathrm{a}}{n_\mathrm{i}} \end{aligned} \tag{2.7b}$$

The total built-in potential across the junction, V_bi, can be found by integrating the field expressions,

$$V_\mathrm{bi} = -\int_{-x_\mathrm{p}}^{x_\mathrm{n}} E_x dx \tag{2.8}$$

However, this just equals the difference in potentials at the junction edges, which allows us to finally obtain

$$V_\mathrm{bi} = V_\mathrm{n} - V_\mathrm{p} = \frac{k_\mathrm{B}T}{q}\ln\frac{N_\mathrm{d}N_\mathrm{a}}{n_\mathrm{i}^2} \tag{2.9}$$

Equation (2.9) can also be obtained in an essentially equivalent manner by noting that the difference between the Fermi levels[10] of the two materials before they are brought together must also equal the potential energy barrier, qV_bi. This can be related to the band edge diagram of the junction (Fig. 2.2b) by noting that the difference in conduction band edges on either side of the junction is also equal to qV_bi:

$$qV_\mathrm{bi} = E_\mathrm{cp} - E_\mathrm{cn} \tag{2.10}$$

Now, using the equations relating the conduction band edge to electron concentration in a semiconductor (see footnote 9), i.e.,

$$\begin{aligned} n_\mathrm{n} &= N_\mathrm{c}e^{-(E_\mathrm{cn}-E_\mathrm{F})/k_\mathrm{B}T} \\ n_\mathrm{p} &= N_\mathrm{c}e^{-\left(E_\mathrm{cp}-E_\mathrm{F}\right)/k_\mathrm{B}T} \end{aligned} \tag{2.11}$$

we can write

$$\frac{n_\mathrm{n}}{n_\mathrm{p}} = \exp\left(\frac{E_\mathrm{cp} - E_\mathrm{cn}}{k_\mathrm{B}T}\right) \tag{2.12}$$

[9] See Appendix A, Sect. A.2.

[10] This is also equivalent to the difference in work functions of the two separated semiconductors (cf. Fig. 2.2a).

Using Eq. (2.10) now gives

$$V_{\mathrm{bi}} = \frac{k_{\mathrm{B}}T}{q}\ln\left(\frac{n_{\mathrm{n}}}{n_{\mathrm{p}}}\right) = \frac{k_{\mathrm{B}}T}{q}\ln\left(\frac{N_{\mathrm{d}}N_{\mathrm{a}}}{n_{\mathrm{i}}^2}\right)$$

as before.

Finally, by using the continuity of $V(x)$ at $x = 0$ in the equations for the potential found above we can write

$$V_{\mathrm{bi}} = V_{\mathrm{n}} - V_{\mathrm{p}} = \frac{q}{2\varepsilon_{\mathrm{s}}}\left(N_{\mathrm{d}}x_{\mathrm{n}}^2 + N_{\mathrm{a}}x_{\mathrm{p}}^2\right) \tag{2.13}$$

which allows us to solve[11] for the depletion widths in terms of the built-in potential:

$$\begin{aligned} x_{\mathrm{n}} &= \left\{\frac{2\varepsilon_{\mathrm{s}}V_{\mathrm{bi}}}{q}\left[\frac{N_{\mathrm{a}}}{N_{\mathrm{d}}(N_{\mathrm{a}}+N_{\mathrm{d}})}\right]\right\}^{1/2} \\ x_{\mathrm{p}} &= \left\{\frac{2\varepsilon_{\mathrm{s}}V_{\mathrm{bi}}}{q}\left[\frac{N_{\mathrm{d}}}{N_{\mathrm{a}}(N_{\mathrm{a}}+N_{\mathrm{d}})}\right]\right\}^{1/2} \end{aligned} \tag{2.14}$$

and thus the total width of the depletion region is given by

$$x_{\mathrm{d}} = x_{\mathrm{n}} + x_{\mathrm{p}} = \left[\frac{2\varepsilon_{\mathrm{s}}}{q}V_{\mathrm{bi}}\left(\frac{1}{N_{\mathrm{a}}} + \frac{1}{N_{\mathrm{d}}}\right)\right]^{1/2} \tag{2.15}$$

Once again we see that the depletion width of a *pn* junction depends most strongly on the material with the lighter doping.

Example 2.1: pn Junction Built-In Potential Calculation A region of n-type silicon with $\rho = 4$ Ω-cm is used to make a *pn* junction with a p-region that has $\rho = 0.2$ Ω cm. Find V_{bi} for the junction.

From the resistivity data for silicon given in Appendix B we can look up that $N_{\mathrm{d}} = 10^{15}$ cm^{-3} and $N_{\mathrm{a}} = 10^{17}$ cm^{-3}. The built-in potential is therefore given by

$$V_{\mathrm{bi}} = \frac{k_{\mathrm{B}}T}{q}\ln\left(\frac{N_{\mathrm{d}}N_{\mathrm{a}}}{n_{\mathrm{i}}^2}\right) \approx 0.7 \text{ V}$$

Panel 2.1: Built-In Potential for Heavily Doped Junctions At very high carrier concentrations due to heavily doped semiconductors the equation derived for V_{bi} above is no longer valid because it is based on the exponential approximation

[11] Using $N_{\mathrm{a}}x_{\mathrm{p}} = N_{\mathrm{d}}x_{\mathrm{n}}$.

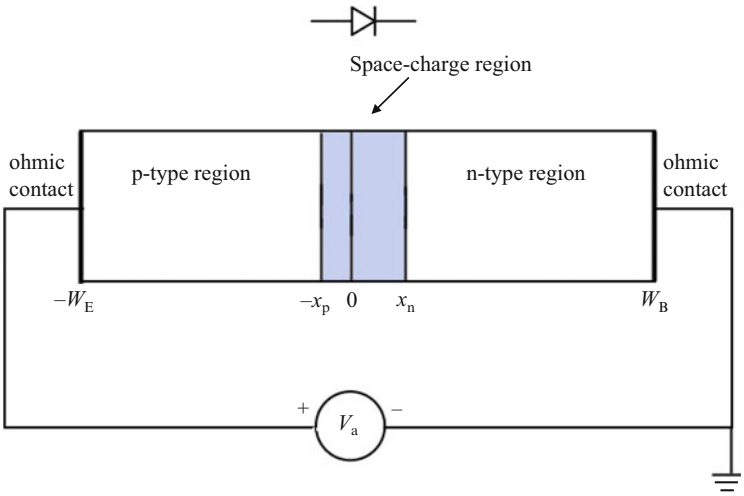


Fig. 2.4 *pn* junction device structure and biasing configuration (Electronic circuit symbol also shown)

of the Fermi–Dirac distribution.[12] When the dopant densities approach N_c or N_v ($\sim 10^{19}$ cm^{-3} for Si) the full carrier distribution function should instead be used for derivations.

However, at very high dopant concentrations the Fermi level lies very near the band edges and the potential (V_n or V_p) in the heavily doped side of the junction is approximately one-half of the band gap energy divided by q or about 0.56 V for silicon. For example, the built-in potential for a *pn* junction composed of heavily doped p-type silicon (denoted p^+n)[13] is

$$V_{bi} = 0.56 \text{ V} + \frac{k_B T}{q} \ln\left(\frac{N_d}{n_i}\right)$$

Similarly, a p^+n^+ junction would have V_{bi} roughly equal to the magnitude of the band gap energy.

2.1.2 *pn* Junction I–V Characteristic

Assume that the *pn* junction is contacted by low-resistance (or ohmic) contacts as illustrated in Fig. 2.4. Since the space-charge region is depleted of mobile carriers, the applied voltage, V_a, will appear almost entirely across this high-resistance potential barrier region of the junction.

If V_a is positive the built-in voltage will be reduced and the junction is said to be *forward biased*, whereas if V_a is negative the built-in voltage is increased and the junction is said to be *reverse biased*. The applied bias will thus alter

[12] See Appendix A, Eq. (A.25).

[13] p^+n and pn^+ junctions are often termed *one-sided pn* junctions.

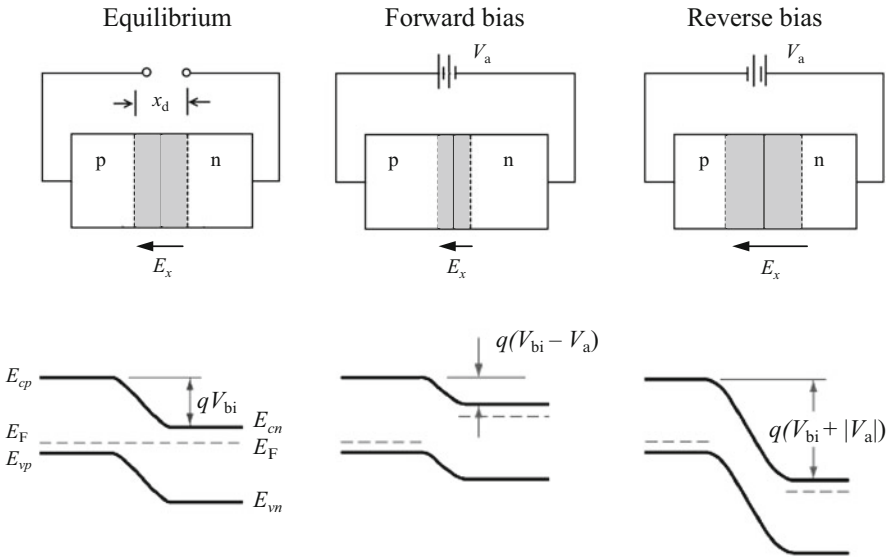


Fig. 2.5 Effect of applied bias on *pn* junction (with similar doping levels on either side). In thermal equilibrium the Fermi level is constant across the entire structure and the built-in electric field opposes the diffusion of carriers across the junction. The applied bias is assumed to drop almost entirely across the depletion layer and thus the Fermi level becomes separated by qV_a. Forward bias reduces the built-in electric field and thus the built-in potential barrier is *reduced* by V_a as shown. Reverse bias on the other hand increases the built-in electric field and the built-in potential is *increased* by V_a

the barrier to diffusion of carriers across the junction that existed in thermal equilibrium as depicted in Fig. 2.5.[14]

Minority carrier densities are crucial to determining current flow through the *pn* junction, whether under forward or reverse bias. We shall see below that the majority carriers act only to supply the minority carrier current "injected" across the junction or to neutralize charge (via recombination) outside of the space-charge region. The reason for this arises from the inherent asymmetry of the *pn* junction; the different doping types lead to few holes in the n-side and conversely few electrons in the p-side. Therefore the majority carriers on either side cannot dominate current flow throughout the *pn* junction structure but must instead facilitate this process from afar, as we now discuss.

In order to derive an expression for the *I–V* characteristic of a *pn* junction we study the dynamics of minority charge carriers in the n- and p-type regions by seeking solutions of the continuity equations for the minority carrier densities on either side of the junction. The most straightforward way of doing this is to assume *low-level injection*, i.e., we assume that the majority carrier concentrations

[14] Note that, more rigorously, the Fermi levels under applied bias should be referred to as *quasi*-Fermi levels since the system is no longer in thermal equilibrium. In this textbook we will not make use of this distinction.

are not appreciably altered by the applied bias. This condition will be seen to be valid for low-to-moderate applied biases, relative to the built-in potential.

Under the assumption of low-level injection the electron density in the n-type region at the boundary with the depletion layer (i.e., at x_n) is equal to the dopant density, N_d, whether at equilibrium or under bias. Similar comments apply to the hole density on the other side of the junction (at $-x_p$). The equilibrium *minority* carrier concentrations on either side of the junction can now be written as [see Eqs. (2.12) and (2.10)]

$$\begin{aligned} n_{p0}(-x_p) &= n_{n0}(x_n)\exp\left(\frac{-qV_{bi}}{k_BT}\right) = N_d(x_n)\exp\left(\frac{-qV_{bi}}{k_BT}\right) \\ p_{n0}(x_n) &= p_{p0}(-x_p)\exp\left(\frac{-qV_{bi}}{k_BT}\right) = N_a(-x_p)\exp\left(\frac{-qV_{bi}}{k_BT}\right) \end{aligned} \tag{2.16}$$

Under an applied bias, V_a, these become

$$\begin{aligned} n_p(-x_p) &= N_d(x_n)\exp\left[\frac{-q(V_{bi}-V_a)}{k_BT}\right] \\ p_n(x_n) &= N_a(-x_p)\exp\left[\frac{-q(V_{bi}-V_a)}{k_BT}\right] \end{aligned} \tag{2.17}$$

We can combine the four equations above to express the *excess* minority carrier concentrations at the junction boundaries in terms of their thermal equilibrium values:

$$\begin{aligned} n'_p(-x_p) &= n_{p0}(-x_p)\left[\exp\left(\frac{qV_a}{k_BT}\right)-1\right] \\ p'_n(x_n) &= p_{n0}(x_n)\left[\exp\left(\frac{qV_a}{k_BT}\right)-1\right] \end{aligned} \tag{2.18}$$[15]

These equations give the important result that the minority carrier density depends exponentially on applied bias while the majority carrier density is assumed insensitive to it (to first order). Note that since the minority carrier densities at thermal equilibrium are typically at least 10 orders of magnitude below the majority carrier densities the assumption of low-level injection will be valid until the exponential factor becomes approximately equal to about 10^{10} for a moderately doped junction and several orders of magnitude larger for a more heavily doped junction.

Having obtained the bias dependence of excess minority carriers at the edges of the depletion region, we can now find solutions of the continuity equations in order to describe how these excess carriers diffuse away from the *pn* junction interface and into the neutral regions, based on an idealized or simplified system, which is typically referred to as *ideal diode analysis*:

[15] Recall form Appendix A, Eq. (A.47), the excess carrier concentration is defined as the difference between the total concentration and the thermal equilibrium concentration.

Since we are interested in the minority carrier diffusion current away from the space-charge region the electric field contribution to this current density is assumed to be negligible.[16] In this case the continuity equations (see Eqs. (A.48) in Appendix A) become

$$\frac{\partial n'}{\partial t} = D_n \frac{\partial^2 n'}{\partial x^2} - \frac{n'}{\tau_n}$$
$$\frac{\partial p'}{\partial t} = D_p \frac{\partial^2 p'}{\partial x^2} - \frac{p'}{\tau_p} \tag{2.19}$$

where it has also been assumed that we have constant doping densities along x (i.e., a step or abrupt junction). Equation (2.19) gives the *diffusion equations*[17] for electrons and holes.

For the case of steady-state (or dc) injection of holes into the n-side of a *pn* junction we can rewrite the diffusion equation as

$$0 = D_p \frac{d^2 p'_{\mathrm{n}}}{dx^2} - \frac{p'_{\mathrm{n}}}{\tau_p} \tag{2.20}$$

This has the exponential solution

$$p'_{\mathrm{n}}(x) = A \ \exp\left(-\frac{x - x_{\mathrm{n}}}{\sqrt{D_p \tau_p}}\right) + B \ \exp\left(\frac{x - x_{\mathrm{n}}}{\sqrt{D_p \tau_p}}\right) \tag{2.21}$$

where A and B are constants and

$$L_p \equiv \sqrt{D_p \tau_p} \tag{2.22}$$

is called the hole *diffusion length*.[18] Similarly, L_n is the corresponding electron diffusion length in the p-type region.

[16] Recall that most of the voltage drop will occur inside the space charge region.

[17] This type of partial differential equation appears in many other fields including heat flow, mass transport in gases and fluids, etc.

[18] The diffusion length represents the average distance a minority carrier travels before recombining.

To evaluate the constants we consider two limiting cases based on the length[19] $x_B = W_B - x_n$ of the neutral n-region from the junction to the ohmic contact:

1. Long-base diode
   If x_B is much greater than the hole diffusion length, essentially all of the injected holes will recombine before reaching the contact. Thus we can neglect the growing exponential term in the solution and set B equal to zero. The constant A can then be found by using Eq. (2.18) for the excess hole concentration at x_n as a function of applied voltage since

$$p'_n(x_n) = A$$

   The solution is therefore

$$p'_n(x) = p_{n0}\left(e^{qV_a/k_BT} - 1\right)\exp\left(-\frac{x - x_n}{L_p}\right) \tag{2.23}$$[20]

   The result is plotted in Fig. 2.6a. The hole diffusion current density can now be found using

$$\begin{aligned} J_p(x) &= -qD_p\frac{dp_n}{dx} \\ &= qD_p\frac{p_{n0}}{L_p}\left(e^{qV_a/k_BT} - 1\right)\exp\left(-\frac{x - x_n}{L_p}\right) \end{aligned} \tag{2.24}$$

   The hole current is therefore greatest at $x = x_n$ (the edge of the space-charge region) and decreases away from the junction because the hole concentration gradient decreases as carriers are lost by recombination. This means that in order for the current to remain constant (as it must in steady state) the *electron current* must *increase* away from the junction as indicated in Fig. 2.6b. This current supplies the electrons with which the holes recombine. On the other hand, at the junction the only electron current flowing is the one injected across the space-charge layer and into the p-region. The electrons injected into the p-region constitute the corresponding minority carrier diffusion current on the other side of the junction.

   Thus, the *total* current flowing through the *pn* junction is obtained by summing the two minority carrier injection currents: holes into the n-side plus electrons into the p-side. The injected electron current can be found by a treatment analogous to

[19] The reason for the subscripts "*B*" and "*E*" in the *pn* junction symbols is for historical reasons relating to the bipolar transistor (Base, Emitter) discussed in Chap. 3.

[20] For simplicity $p_{n0}(x_n)$ is written as p_{n0}. Note that for a step junction p_{n0} will be constant throughout the neutral n-type region.

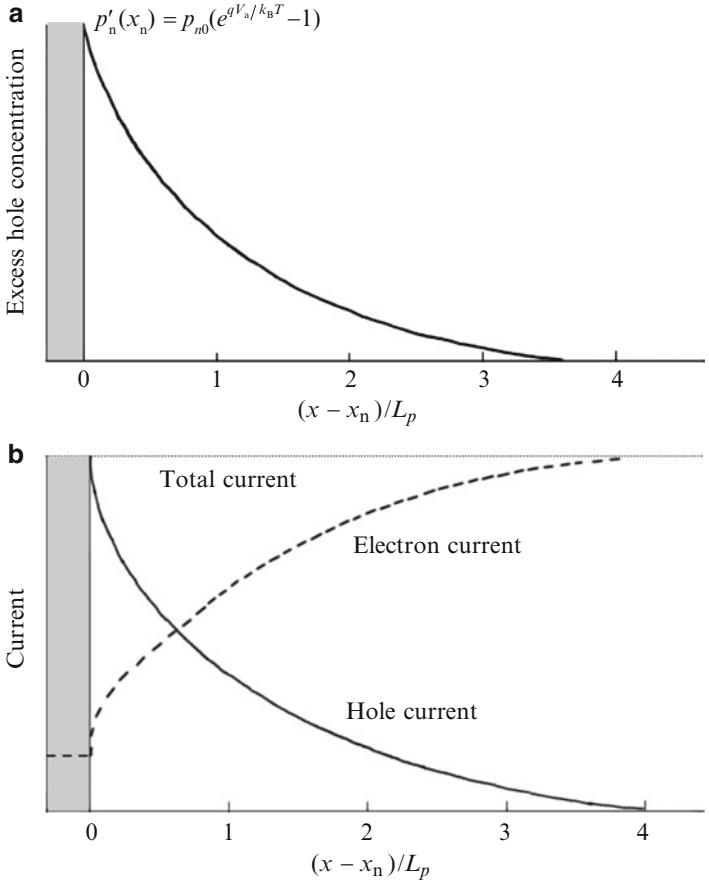


Fig. 2.6 Long-base diode under forward bias. (**a**) Excess minority hole concentration vs. distance in the n-region. (**b**) Electron and hole currents in the n-region

that used above for holes (still under the assumption of a long-"base" diode, so that $x_E = W_E - x_p$ is much greater than the electron diffusion length):

$$J_n(x) = qD_n \frac{n_{p0}}{L_n}\left(e^{qV_a/k_BT} - 1\right)\exp\left(\frac{x + x_p}{L_n}\right) \qquad (2.25)^{21}$$

To obtain an expression for the total current we now sum the minority carrier current components at $-x_p$ and $+x_n$ to get

[21] Note that the x-coordinate is negative in this expression.

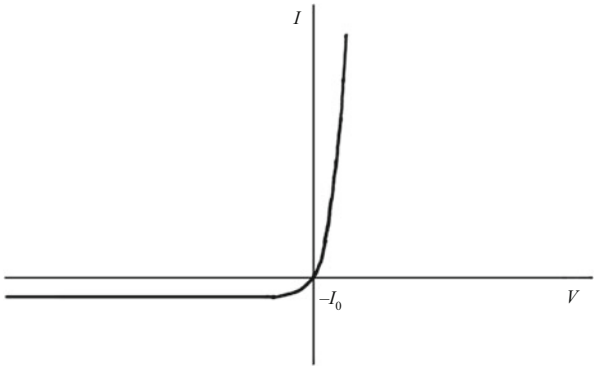


Fig. 2.7 Ideal *pn* junction *I–V* characteristic

$$J = J_p(x_\mathrm{n}) + J_n(-x_\mathrm{p}) = qn_\mathrm{i}^2\left(\frac{D_p}{N_\mathrm{d}L_p} + \frac{D_n}{N_\mathrm{a}L_n}\right)\left(e^{qV_\mathrm{a}/k_\mathrm{B}T} - 1\right)$$
$$= J_0\left(e^{qV_\mathrm{a}/k_\mathrm{B}T} - 1\right) \tag{2.26}$$

where J_0 is the magnitude of the (reverse) *saturation current density* predicted by this model for negative applied bias greater than a few $k_\mathrm{B}T/q$ volts (expressed above for the case of a step *pn* junction). This expression is known as the *ideal diode equation* and is plotted in Fig. 2.7. This form of equation and the exponential dependence is very common in semiconductor electronic devices since it ultimately derives from the fundamental carrier statistics inside these materials, which are governed by the Fermi–Dirac (or Maxwell–Boltzmann) distributions.[22]

2. Short-base diode
   In this case the lengths x_B and x_E of the n- and p-type regions are much shorter than the diffusion lengths L_p and L_n. Thus very little recombination occurs in the bulk of the n- and p-type regions. In this limit, almost all the injected minority carriers recombine at the ohmic contacts at either end of the diode structure and thus the excess minority carrier distribution can be approximated as being essentially linear, so that

$$p'_\mathrm{n}(x) = A' + B'\frac{x - x_\mathrm{n}}{L_p} \tag{2.27}$$

To find the constants we once again apply appropriate boundary conditions. The ohmic contact at $x = W_\mathrm{B}$ can be considered a perfect sink for any excess carriers and therefore

[22] The applied bias can be thought of as a small perturbation to the equilibrium properties of a material; hence, even when current flows through electronic devices they are still close to thermal equilibrium under most conditions.

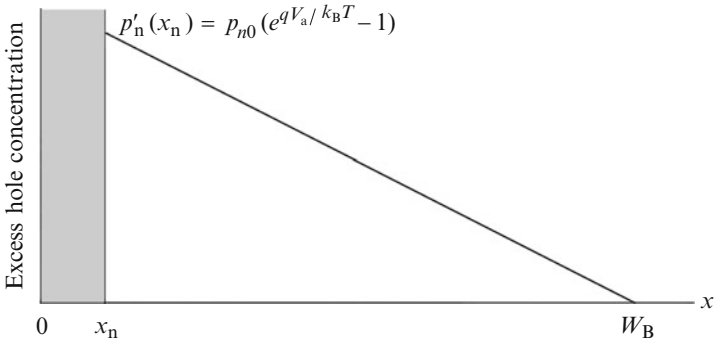


Fig. 2.8 Short-base diode excess hole carrier distribution in n-region under forward bias

$$p'_{\mathrm{n}}(W_{\mathrm{B}}) = 0$$

The other boundary condition is that

$$p'_{\mathrm{n}}(x_{\mathrm{n}}) = A'$$

as in the long-base diode.

The solution for the excess hole density in the n-type region therefore becomes

$$p'_{\mathrm{n}}(x) = p_{\mathrm{n0}}\left(e^{qV_{\mathrm{a}}/k_{\mathrm{B}}T} - 1\right)\left(1 - \frac{x - x_{\mathrm{n}}}{x_{\mathrm{B}}}\right) \tag{2.28}$$

and is plotted in Fig. 2.8. Note the assumption that no recombination occurs in the n-type region is equivalent to letting the lifetime τ_p approach infinity in the diffusion equation. The differential equation that results has a linear solution. A linearly varying concentration indicates that the hole current remains constant throughout the n-type region and that no electron current is needed to compensate for recombining holes.

We can now calculate the hole diffusion current as before

$$J_p(x) = -qD_p\frac{dp_{\mathrm{n}}}{dx} = qD_p\frac{p_{\mathrm{n0}}}{x_{\mathrm{B}}}\left(e^{qV_{\mathrm{a}}/k_{\mathrm{B}}T} - 1\right) \tag{2.29}$$

and also the total current flowing through the thin diode by summing the injected hole and electron diffusion currents to get

$$J = qn_{\mathrm{i}}^2\left(\frac{D_p}{N_{\mathrm{d}}x_{\mathrm{B}}} + \frac{D_n}{N_{\mathrm{a}}x_{\mathrm{E}}}\right)\left(e^{qV_{\mathrm{a}}/k_{\mathrm{B}}T} - 1\right) \tag{2.30}$$

This once again has the form of the ideal diode equation.

We see that the currents through the short-base and long-base diode, Eqs. (2.23) and (2.30), are very similar except for the characteristic length associated with

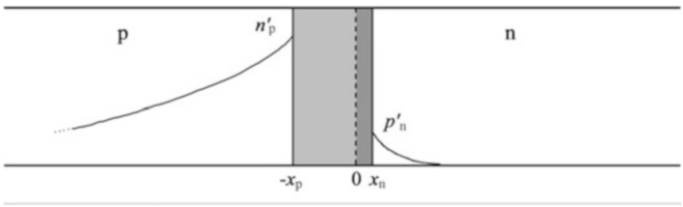


Fig. E2.2 Sketch of excess minority carrier concentrations under forward bias (not to scale). Note the relative magnitudes of the minority carrier concentrations in addition to the difference in diffusion lengths for this problem

each geometry—the minority carrier diffusion length in the long-base diode case and the width of the p- or n-type region in the short-base diode case.

A given diode can be approximated by a combination of these two limiting cases; for example, it may be short in the p-type region and long in the n-type region, or vice versa. In these cases we simply take the proper combinations of the minority current densities derived above to come up with the total junction current.[23]

Example 2.2: Ideal Diode Equation Current Components Calculate the hole and electron contributions to the total current density in a forward-biased (ideal, long-base diode) *pn* junction. The diode is made from silicon with n-side doping $N_d = 10^{18}\ \text{cm}^{-3}$ and p-side doping $N_a = 5 \times 10^{16}\ \text{cm}^{-3}$. Assume that $\tau_p = 1\ \mu\text{s}$, $\tau_n = 10\ \mu\text{s}$, and an applied bias of 0.7 V. Sketch the excess carrier concentration on either side of the depletion region.

The long-base ideal diode equation currents are given by Eq. (2.26):

$$J_p(x_n) = \frac{qn_i^2 D_p}{N_d L_p}\left(e^{qV_a/k_BT} - 1\right); J_n(-x_p) = \frac{qn_i^2 D_n}{N_a L_n}\left(e^{qV_a/k_BT} - 1\right)$$

Using the mobility data for silicon as a function of impurity concentration (Appendix B) the diffusion coefficients can be found, which leads to diffusion lengths of approximately 20 μm and 150 μm for holes and electrons, respectively. Substituting these values along with the data given in the problem results in

$$J_p(x_n) \approx 35\ \text{mA/cm}^2; J_n(-x_p) \approx 515\ \text{mA/cm}^2$$

A sketch of the excess carrier concentrations is shown in Fig. E2.2. The ratio of hole or electron current to the total current flowing through the junction is known as the *injection efficiency* and will be seen in Chap. 3 to be a key factor in the performance of bipolar transistors.

[23] If one/both of the diode regions have intermediate lengths (i.e., x_B or x_E is comparable to L_p or L_n, respectively), in other words they are neither short nor long, then the diffusion equations lead to solutions for the excess carrier concentrations containing hyperbolic functions [1]. The net result once more is a modification of the saturation current in the ideal diode equation.

In summary, the diode equations predict a large current flow under forward bias and a small constant saturation current under reverse bias. This asymmetry resulted because forward bias aids the injection of carriers from each region across the junction. Under reverse bias the net flow across the junction is instead composed of minority carriers from each region (i.e., electrons on the p-side and holes on the n-side). These are few in number and hence only a small current flows under reverse bias. This implies that the more lightly doped side of the junction determines most of the reverse saturation current, which can also be seen from the diode equations. For example, if the doping on the *n*-side is much less that that on the *p*-side, the hole flow under reverse bias across the junction into the *p*-side will be much greater than the corresponding electron flow into the *n*-side. Since minority carriers on each side of the space-charge region are swept away by the large electric field present in the junction under reverse bias, their concentration is reduced below thermal equilibrium values. This leads to thermal generation of minority carriers in the vicinity of the junction, which is what supplies the carriers for the reverse saturation current in an ideal diode.

Lastly, we note in this section that the depletion width as a function of applied bias can be found in a straightforward manner by modifying Eq. (2.15) for an abrupt *pn* junction by replacing the built-in potential with (V_{bi}–V_a):

$$x_d = x_n + x_p = \left[\frac{2\varepsilon_s}{q}\left(\frac{1}{N_a}+\frac{1}{N_d}\right)(V_{bi}-V_a)\right]^{1/2} \tag{2.31}$$

We have seen that if V_a is positive, the built-in barrier at the junction will be reduced. In addition, Eq. (2.31) tells us that the depletion region also becomes *narrower* (see Fig. 2.5). Conceptually, we can think of the applied voltage as moving majority carriers toward the edges of the depletion region where they neutralize some of the space charge, which reduces the overall depletion region width. Similarly, if a negative voltage is applied across the junction, the built-in barrier is increased and majority carriers are pulled away from the edges of the depletion region, which therefore *widens*. When the magnitude of V_a becomes considerably larger than V_{bi}, the depletion width varies as the square root of reverse bias.

Similarly, the magnitude of the maximum electric field at the junction, E_{max}, and its relationship to applied voltage can be found by noting that the area under the field curve represents the potential across the junction. We saw earlier that for a step junction the field varies linearly with distance (Fig. 2.3c) and therefore in thermal equilibrium we get

$$\frac{1}{2}E_{max}x_d = V_{bi} \Rightarrow E_{max} = \frac{2V_{bi}}{x_d} \tag{2.32a}$$

which is equivalent to the maximum field obtained using Eqs. (2.3) and (2.4). Under an applied bias this becomes

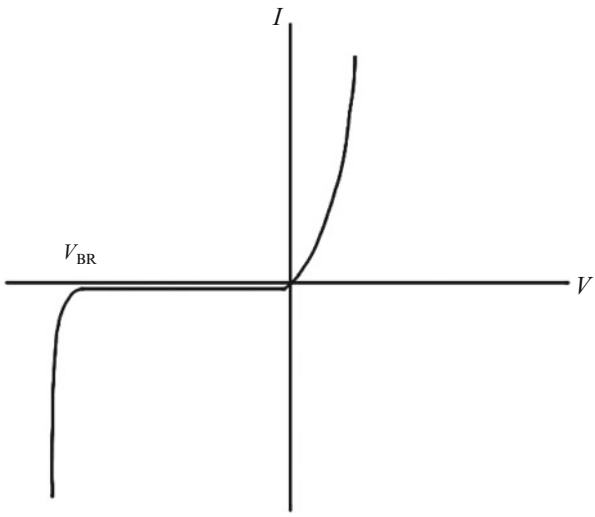


Fig. 2.9 *pn* junction *I–V* characteristic illustrating reverse-bias breakdown

$$E_{\max} = \frac{2(V_{\mathrm{bi}} - V_{\mathrm{a}})}{x_{\mathrm{d}}} \tag{2.32b}$$

and by using Eq. (2.31) the explicit dependence of the maximum electric field on applied bias can be found.[24]

2.1.3 Deviations from Ideal Behavior

1. Reverse-bias breakdown
   An effect not accounted for in the treatment of the ideal diode above is *electrical breakdown* of the junction under large reverse bias, as illustrated schematically in Fig. 2.9. As the reverse bias voltage is increased a critical value is eventually reached where the resulting electric field in the space-charge region causes the magnitude of the current to increase sharply. Generally speaking, such breakdown does not physically damage the junction[25] and is an important process that helps define its stable *I–V* characteristic and potential applications.

   There are two basic mechanisms[26] that cause junction breakdown under high electric fields:

[24] It can be seen that $E_{\max} \propto V_{\mathrm{a}}^{1/2}$ for large reverse biases.

[25] One must take care however to avoid mechanical and/or thermal breakdown (in particular *thermal runaway* processes that cause uncontrollable positive current feedback loops), which can lead to device failure.

[26] The breakdown phenomenon of *punchthrough* may also occur for short-base diodes under reverse bias. This effect is discussed in Sect. 3.4.

– Avalanche breakdown
Consider an electron traveling in the space-charge region of a reverse-biased *pn* junction: The electron travels an average distance λ, its mean free path, before interacting with an atom in the lattice and losing energy by scattering. The energy gained by the electron between collisions is given by the work done on it by the electric field, which can be written as

$$\Delta \mathrm{E} = q\int_0^{\lambda} \vec{E} \cdot d\vec{x} \tag{2.33}$$

If the electron gains sufficient energy[27] from the field before colliding with an atom in the lattice it can excite an electron out of the silicon–silicon (or other semiconductor) bond so that *three carriers*—the initial electron and an additional electron–hole pair—are created after the collision. This process is illustrated in Fig. 2.10a. Each of the three carriers can then cause similar collisions, etc., which leads to a sudden multiplication or avalanching of the number of charge carriers available to participate in current flow.[28] Hence, current through the junction increases very rapidly.

The multiplication factor for avalanching current compared to the ideal diode saturation current can be written

$$M \equiv \frac{|I|}{I_0} \tag{2.34}$$

and is illustrated in Fig. 2.10b. Empirically, the avalanching multiplication factor for a reverse-biased *pn* junction can be written as

$$M \equiv \frac{1}{1 - (|V_\mathrm{a}|/V_\mathrm{BR})^m} \tag{2.35}$$

where V_BR is the breakdown voltage (taken to be the point at which the current increases rapidly), and m is observed to vary between 2 and 6. Qualitatively, the breakdown voltage will increase with increasing band gap energy (since more energy is required to excite the e–h pairs); decrease with increasing doping level[29] [since the maximum electric field will increase—Eqs (2.3) and (2.4)]; and increase with increasing temperature [due to the increased lattice scattering which limits the mean free path in Eq. (2.33)].

[27] Note that this energy must be at least on the order of the band gap in order to create a new electron–hole pair.

[28] This process is more generally referred to as *impact ionization*.

[29] At very large doping levels the mean free path will also decrease significantly due to impurity scattering; however, in this regime avalanche breakdown is less likely to occur.

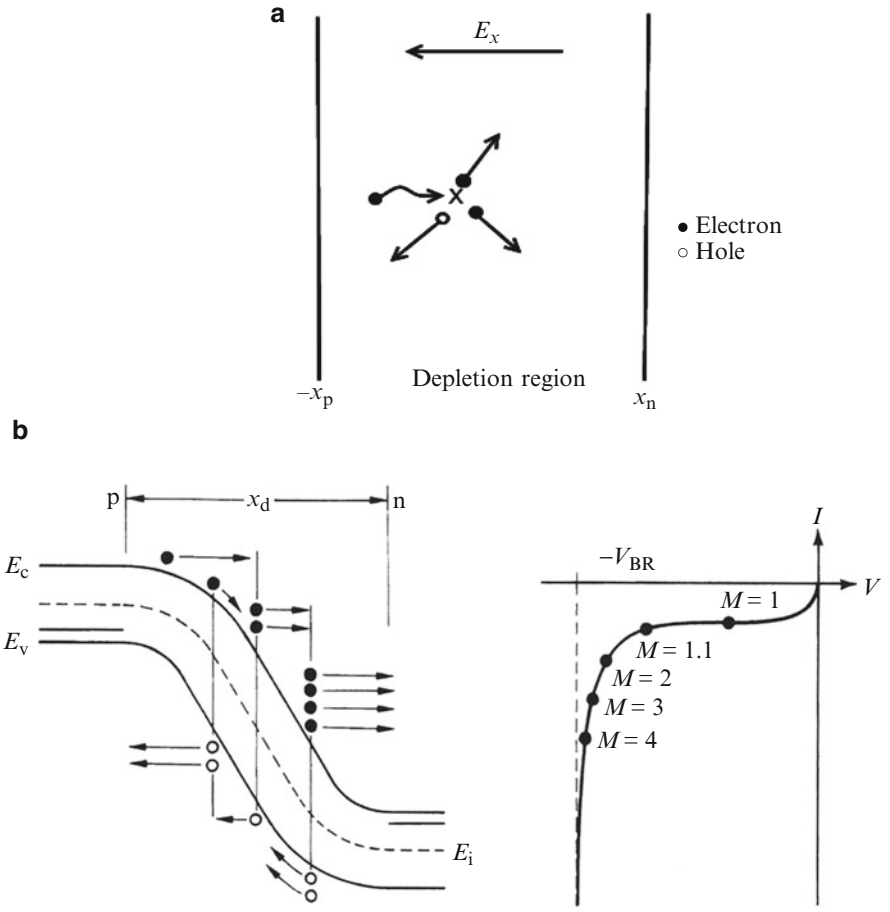


Fig. 2.10 (**a**) Illustration of avalanche breakdown process. The incoming electron (*wavy arrow*) has gained sufficient energy from the field inside the space-charge region such that when it collides with the lattice it is able to free an electron from its bond resulting in the creation of a new electron–hole pair. (**b**) Avalanche breakdown process superimposed on *pn* junction band edge diagram and *I*–*V* curve for reverse bias showing increasing multiplication factor near breakdown voltage (After [7])

– Zener breakdown
Recall that the width of the depletion region decreases as the dopant concentration increases:

$$x_d = x_n + x_p = \left[\frac{2\varepsilon_s}{q}\left(\frac{1}{N_a} + \frac{1}{N_d}\right)V_{bi}\right]^{1/2} \tag{2.36}$$

Thus the depletion region can be quite narrow for high doping levels resulting in the energy bands across the junction region being bent more steeply. When the width of the depletion layer is small enough, *tunneling*

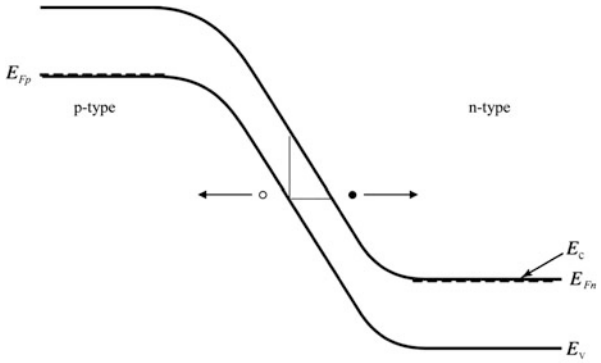


Fig. 2.11 Zener tunneling in a *pn* junction under reverse bias. For heavy doping the narrow depletion width allows an electron from the valence band to tunnel through the band gap leaving behind an empty state (*hole*). The triangular tunneling barrier (*grey lines*) has a height equal to the band gap and width that is inversely proportional to the slope of the band edges (or the electric field strength) (Adapted from [4])

through the band gap can occur (Fig. 2.11). This type of tunneling of electrons from the valence band to the conduction band is called Zener tunneling.[30] Recall[31] that the probability for a particle to tunnel through a barrier is a strong function of the thickness of the barrier and so Zener tunneling is only significant in heavily doped junctions, in which the fields are high and the depletion region is narrow. If we decrease the dopant concentrations, the width of the space-charge region increases and the probability of tunneling decreases rapidly—avalanche breakdown then becomes more likely than Zener breakdown. Zener breakdown also has the opposite temperature dependence compared to avalanching: the tunneling breakdown voltage *decreases* as temperature is increased since the average thermal energy of the electrons will be greater, thus allowing them to tunnel through the barrier more easily, and the band gap energy (or barrier height) will also decrease with temperature.[32]

Devices exhibiting Zener breakdown generally have lower breakdown voltages than those that break down by avalanching. For example, in silicon pure Zener breakdown is usually found in diodes having a breakdown voltage of less than 5 V (Fig. 2.12). As the breakdown voltage increases there will be

[30] Such field-induced *interband tunneling* can occur more generally in solids (e.g., insulators) and is often referred to simply as Zener breakdown; C. M. Zener, Proc. Roy. Soc. (London) **A145**, 523 (1934).

[31] See Appendix A, Example A.4.

[32] See Appendix B for the band gap energy as a function of temperature. Increased thermal energy results in larger crystal lattice spacing, which usually leads to the band structure resembling more of a free particle (i.e., a smaller band gap).

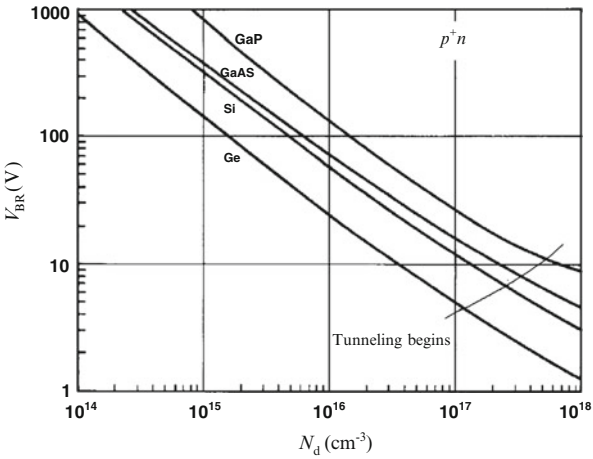


Fig. 2.12 Reverse-bias breakdown data for a one-sided step *pn* junction (Adapted from S. M. Sze, G. Gibbons, Appl. Phys. Lett. **8**, 111 (1966))

a transition region in which both avalanche and tunnel breakdown can occur simultaneously.

Commercially available diodes have well-defined breakdown characteristics and are generally referred to as Zener diodes regardless of their particular breakdown mechanism. Since the voltage across a diode at breakdown is essentially independent of current, Zener diodes are often used as voltage regulators in circuits that require a known value of voltage.

2. Space-charge generation and recombination currents

The analysis used to derive the ideal diode equation ignored processes that occur inside the depletion region, which alter the observed *I–V* characteristic of real *pn* junctions over certain bias ranges compared to the ideal case. Previously we treated the space-charge region simply as a barrier to the diffusion of majority carriers across the junction. However, like the rest of the diode structure, the space-charge region also contains generation–recombination centers.[33] Under forward bias, the injected carriers must pass through space-charge region and a portion of these will recombine before reaching the neutral n- or p-type regions (at x_n or $-x_p$). On the other hand, under reverse bias the opposite process will occur as generation of carriers in the depleted space-charge region also leads to additional current above that predicted by the ideal diode equation analysis. In other words, the ideal diode analysis, in particular its boundary conditions and resulting currents, is still valid; however, the additional current components due to generation/recombination occurring in the space-charge region were not accounted for and must therefore be added to the ideal current we found above.

To find expressions for generation–recombination currents in the space-charge region we may use Shockley–Hall–Read (SHR) (recombination) theory.

[33] Or, more generally, generation/recombination can occur in the space-charge region through a variety of mechanisms.

The essential result of such a treatment is that the space-charge recombination current under an appreciable forward bias can be expressed as

$$J_{\mathrm{R}} \approx \frac{q x_{\mathrm{d}} n_{\mathrm{i}}}{2\tau} \exp\left(\frac{qV_{\mathrm{a}}}{2k_{\mathrm{B}}T}\right) \tag{2.37}$$[34]

where τ is the carrier lifetime in the space-charge region. If τ is assumed to be approximately equal to the electron and hole lifetimes outside the space-charge region, the ratio between the ideal forward-bias diode current [long-base diode case; Eq. (2.26)] and the space-charge recombination current within this model is

$$\frac{J_{\mathrm{ideal}}}{J_{\mathrm{R}}} \approx \frac{2n_{\mathrm{i}}}{x_{\mathrm{d}}} \left[\frac{L_n}{N_{\mathrm{a}}} + \frac{L_p}{N_{\mathrm{d}}}\right] \exp\left(\frac{qV_{\mathrm{a}}}{2k_{\mathrm{B}}T}\right) \tag{2.38}$$

The space-charge recombination current will therefore become less significant relative to the ideal diode current as forward bias increases. However, the different exponential behavior of J_{R} can be observed in real *pn* diodes, especially at low currents. For typical silicon diodes, J_{ideal} exceeds J_{R} for applied biases greater than about 0.3–0.4 V. In addition, since the ratio is proportional to the intrinsic carrier concentration, *pn* junctions based on larger band gap semiconductors (smaller n_{i}) will be affected more strongly by recombination processes inside the space-charge region.

Under reverse bias, the net generation of carriers in the space-charge region results in a generation current, J_{G}, that is proportional to the width of the depletion layer, x_{d}, and analogous to Eq. (2.38) we can write the ratio as

$$\frac{J_{\mathrm{ideal}}}{J_{\mathrm{G}}} \approx \frac{2n_{\mathrm{i}}}{x_{\mathrm{d}}} \left[\frac{L_n}{N_{\mathrm{a}}} + \frac{L_p}{N_{\mathrm{d}}}\right] \tag{2.39}$$

Practical values of the parameters for silicon *pn* diodes are such that $J_{\mathrm{ideal}}/J_{\mathrm{G}}$ is usually much less than unity and thus the current in reverse-biased silicon *pn* diodes is generated primarily inside the space-charge region as opposed to outside (as in the ideal diode). This behavior once again is more pronounced as the semiconductor band gap energy increases. Since the width of the depletion layer will vary as the square root of applied reverse bias (Eq. 2.31), the generation current in the space-charge region is also a weak function of reverse bias and will gradually increase instead of being the constant reverse saturation

[34] The recombination current density in the space-charge region can be found by evaluating the expression

$$J_{\mathrm{R}} = q\int_{-x_{\mathrm{p}}}^{x_{\mathrm{n}}} U dx,$$

where U is the recombination rate (in units of $\mathrm{s}^{-1}\ \mathrm{cm}^{-3}$). For SHR theory U can be approximated by Eq. (A.49c) in Appendix A, and by using Eq. (2.44) the integral above can be evaluated giving the result of Eq. (2.37). Note that in more detailed recombination models (and in practice) the factor multiplying $k_{\mathrm{B}}T$ in the exponential can differ from 2.

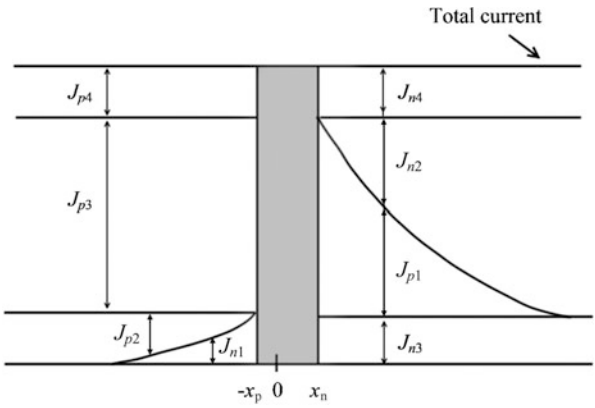


Fig. 2.13 *pn* junction forward-bias current components including space-charge region contributions. Components 1–3 are based on the ideal diode analysis, while the remaining contribution (4) comes from electrons and holes recombining in the space-charge layer shown in gray (Adapted from [4])

current predicted by the ideal diode equation. Lastly, note that ideal diode behavior will become more dominant in both forward and reverse biases as temperature is increased due to the increase in intrinsic carrier concentration.

The observed steady-state dependence of total *pn* junction current on voltage is obtained by summing the ideal diode current and the space-charge generation–recombination current components:

$$I = I_0\left[\exp\left(\frac{qV_a}{k_BT}\right) - 1\right] + I_{GR0}\left[\exp\left(\frac{qV_a}{2k_BT}\right) - 1\right] \qquad (2.40)^{35}$$

We now have a more complete picture of the currents flowing across a *pn* junction as shown in Fig. 2.13.

3. Effect of series resistance

An initial assumption we made in deriving Eqs. (2.26) and (2.30) was that the applied bias, V_a, is dropped entirely across the junction space-charge region. By examining some typical diode structures (see, e.g., Problem 2) it is possible to get a feel for the accuracy of this assumption. In most cases, it is reasonable to conclude that we can safely assume that the entire applied voltage changes the height of the potential barrier at the *pn* junction for *low-to-moderate* current densities. However, if the applied forward bias is nearly as large as the built-in potential, the barrier preventing carriers from diffusing across the junction is substantially reduced and large currents can flow. Essentially, the high-resistance depletion layer is almost eliminated in this case (the band edges are flattened) and a significant fraction of the applied voltage is dropped across the neutral regions in series with the junction. (This is the reason that the actual forward voltage applied across the *pn* junction is never as large as the built-in voltage.)

[35] I_{GR0} is found from Eq. (2.37).

The effect of series resistance can be included by modifying the ideal diode equation as follows:

$$I = I_0\left[\exp\left(\frac{q(V_a - IR_s)}{k_B T}\right) - 1\right] \tag{2.41}$$

where R_s is the effective (parasitic) series resistance.[36]

In general, the diode *I–V* relationship may be written in an empirical parametric form as

$$I = I'_0\left(e^{qV_a/\eta k_B T} - 1\right) \tag{2.42}$$[37]

where I'_0 is now the effective saturation current and η is called the *ideality factor*. We have seen that the ideality factor will approach 2, when recombination processes in the space-charge region are important, and generally will have a value between 1 (ideal diode) and 2. For applied voltages greater than a few $k_B T/q$ the -1 term can be neglected and taking the natural log of both sides yields

$$\ln I = \ln I'_0 + \frac{qV_a}{\eta k_B T} \tag{2.43}$$

Using a semi-log scale we can now describe the overall diode forward *I–V* characteristic including non-idealities as shown in Fig. 2.14.

An additional non-ideality at large forward bias indicated in Fig. 2.14 is the effect of *high-level injection*. As mentioned earlier, our assumption of low-level injection used to derive the ideal diode equation will begin to breakdown once the forward bias reaches a level that causes the excess minority carrier concentrations to be comparable to the majority carrier or doping levels of the *pn* junction neutral regions. In this case, excess *majority* carriers that act to maintain space-charge neutrality become significant and we can no longer assume that the majority carrier concentration is equal to the dopant density. In order to get a feel for the underlying behavior in the condition of high-level injection, let us examine the hole–electron product at the edges of the depletion region using Eqs. (2.17):

$$p_n(x_n)n_n(x_n) = p_p(-x_p)n_p(-x_p) = n_i^2\exp\left(\frac{qV_a}{k_B T}\right) \tag{2.44}$$

This equation is similar in form to the mass-action law for semiconductors in thermal equilibrium, except that the constant product has a greater value. If we

[36] Contact resistances may also be lumped into this parameter.

[37] The effects of series resistance may also be explicitly included here as in Eq. (2.41).

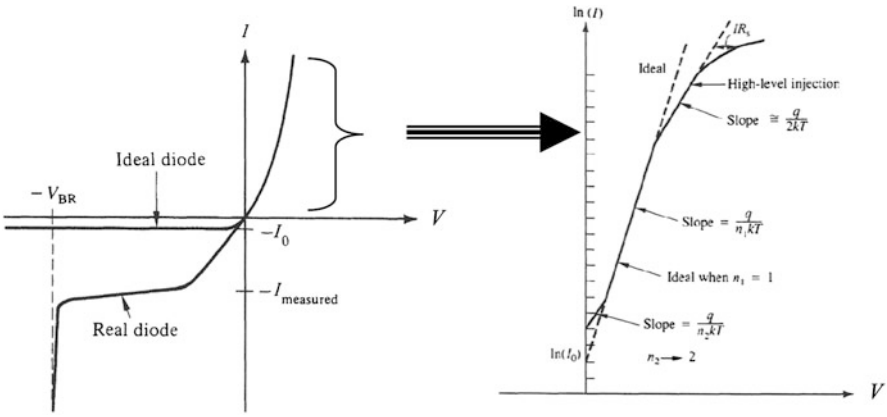


Fig. 2.14 *pn* junction *I*–*V* characteristic including non-idealities (After [7])

now make the conjecture that Eq. (2.44) holds generally,[38] then this *pn* product will be obeyed regardless of low- or high-level injection conditions. In the case of high-level injection this implies that if the concentration of minority carriers approaches the majority carrier concentration, i.e., *p* and *n* become similar in magnitude, then

$$p_{\mathrm{n}}(x_{\mathrm{n}}) \approx n_{\mathrm{i}} \exp\left(\frac{qV_{\mathrm{a}}}{2k_{\mathrm{B}}T}\right) \tag{2.45}$$

with a similar result for electrons in the p-side. This is the basis for the ideality factor approaching 2 at large biases in Fig. 2.14. Note that the effects of series resistance and high-level injection may occur simultaneously in a given *pn* diode device and care must be taken when analyzing the large forward bias portion of the *I*–*V* characteristic.

The next step in analyzing *pn* junctions would be to consider non-uniform dopant densities instead of the abrupt or step junctions we have been considering thus far. For example, in a *linearly graded junction*[39] the dopants have a linear concentration profile from the p- to n-type material instead of being constant. Using the depletion approximation and electrostatics allows one to find a linear charge dependence in the resulting space-charge region and thus a field and potential that vary quadratically and to the third power, respectively.

[38] In other words, we are assuming that *quasi-equilibrium* holds. It can be shown more generally that Eq. (2.44), often referred to as the "law of the junction", is also valid inside the space-charge region and that $pn \leq n_{\mathrm{i}}^2 e^{qV_{\mathrm{a}}/k_{\mathrm{B}}T}$ throughout the *pn* junction (H. K. Gummel, Solid-State Electron. **10**, 209 (1967)).

[39] This can sometimes be used to approximately model *pn* junctions formed via gas phase diffusion processes. (An exponential distribution may also be used.)

Most junctions encountered in practice can be usually be approximated by either a step junction or other analytical profiles (sometimes dependent on the bias applied), which allows important physical insight into their properties. Arbitrary doping profiles are usually handled numerically if a very precise result is required. Similar comments apply to the currents flowing through such junctions. In Chap. 3 it will be seen that approximations can be made that, not surprisingly, show that the ideal diode law is still obtained in practical cases of nonuniform doping although the expression for reverse saturation current will in general be different.

2.1.4 *Small-Signal Parameters*[40]

Small amplitude time-varying signals are often applied to electronic devices that are operating at a given dc bias point.[41] The response of a device to these small perturbations is characterized by a set of small-signal parameters. This type of approach is particularly useful for applications in amplification, communications, and signal processing where the overall nonlinear device behavior can be approximated by linear components (the small-signal parameters) that are valid for small excursions from the dc biasing point.

2.1.4.1 Conductance

The small-signal or differential conductance of the *pn* junction is given by

$$G = \frac{dI}{dV_a} = \frac{q}{k_B T} I_0 \left(e^{qV_a/k_B T} \right) = \frac{q}{k_B T} (I + I_0) \tag{2.46}$$[42]

The small-signal conductance of a *pn* junction is thus seen to be very sensitive and directly proportional to the current under appreciable forward bias, while it becomes zero under appreciable reverse bias for the idealized junction.

2.1.4.2 Diffusion Capacitance

The variation of minority carrier charges stored in the neutral regions of the *pn* junction under forward bias contributes a small-signal capacitance known as

[40] The results of this section are based on an ideal step *pn* junction diode unless otherwise noted. Non-idealities and/or other junction doping profiles can be included in a straightforward manner.

[41] This includes a device that is unbiased (i.e., zero bias).

[42] Not to be confused with the same symbol also used for carrier generation rate.

the diffusion capacitance, C_{d}, which characterizes how quickly the minority carrier charge distribution can change in response to a time-varying voltage. The excess minority carrier charge per unit area, Q_n or Q_p, can be used to find the small-signal diffusion capacitance (per unit area) using $C_{\mathrm{d}} = |dQ_{n,p}/dV_{\mathrm{a}}|$. For the short-base diode the minority carrier charge can be found using Eq. (2.28)[43] giving

$$C_{\mathrm{d}} = \frac{q}{k_{\mathrm{B}}T}\left(\frac{qx_{\mathrm{B}}p_{\mathrm{n0}}}{2} + \frac{qx_{\mathrm{E}}n_{\mathrm{p0}}}{2}\right)e^{qV_{\mathrm{a}}/k_{\mathrm{B}}T} \tag{2.47}$$

where the contribution from both minority electrons and holes on either side of the junction has been included. In the case of a $p^{+}n$ diode the diffusion capacitance can be expressed as

$$C_{\mathrm{d}} \approx G\tau_{\mathrm{t}} \tag{2.48}$$

where a characteristic time, τ_{t}, multiplies the conductance. To see the significance of this term note that

$$\tau_{\mathrm{t}} = \frac{x_{\mathrm{B}}^2}{2D_p} = \frac{Q_p}{J_p} \tag{2.49}$$

The amount of stored charge divided by the rate at which the charge enters or leaves the n-type region is equal to the average time a carrier spends in this region and we call this the average *transit time* of a hole moving through the n-region of the short diode. Similar comments apply to electrons moving through the p-region of the diode.

The long-base diode diffusion capacitance has to take into account the recombination of the minority charges, which normally requires solving the time-dependent continuity or diffusion equations. However, we noted earlier that the long- and short-base diode equations are identical if the diffusion lengths are interchanged with the widths of the neutral regions. Therefore Eq. (2.47) can be modified to give the long-base diode diffusion capacitance as follows:

$$C_{\mathrm{d}} = \frac{q}{k_{\mathrm{B}}T}\left(\frac{qL_pp_{\mathrm{n0}}}{2} + \frac{qL_nn_{\mathrm{p0}}}{2}\right)e^{qV_{\mathrm{a}}/k_{\mathrm{B}}T} \tag{2.50}$$

[43] This is simply the area of the triangle representing the excess minority carrier distribution *vs.* distance.

In the case of a p^+n diode this simplifies to

$$C_\mathrm{d} \approx \frac{G\tau_p}{2} \tag{2.51}$$[44]

and vice versa for a pn^+ diode.

As expected, the diffusion capacitance is negligible under reverse bias because the minority carrier storage is small.

2.1.4.3 Junction Capacitance

In addition to the diffusion capacitance due to excess minority carriers, the fixed charges inside the depletion layer will contribute what is known as the depletion or junction capacitance, C_j, which characterizes the expansion or contraction of majority carrier distributions near the edges of the depletion region in response to a time-varying bias. The space charge per unit area on either side of the (metallurgical) junction is given by

$$Q_s = qN_\mathrm{d}x_\mathrm{n} = qN_\mathrm{a}x_\mathrm{p} \tag{2.52}$$

which can be used to find the small-signal capacitance C_j (per unit area) of the junction:

$$C_\mathrm{j} = \left|\frac{dQ_s}{dV_\mathrm{a}}\right| = qN_\mathrm{d}\frac{dx_\mathrm{n}}{dV_\mathrm{a}} = qN_\mathrm{a}\frac{dx_\mathrm{p}}{dV_\mathrm{a}} \tag{2.53}$$

Since $x_\mathrm{p} = (N_\mathrm{d}/N_\mathrm{a})x_\mathrm{n}$ and $x_\mathrm{d} = x_\mathrm{n} + x_\mathrm{p}$ we can write

$$x_\mathrm{d} = x_\mathrm{n}(1 + N_\mathrm{d}/N_\mathrm{a}) = \left[\frac{2\varepsilon_\mathrm{s}}{q}\left(\frac{1}{N_\mathrm{a}} + \frac{1}{N_\mathrm{d}}\right)(V_\mathrm{bi} - V_\mathrm{a})\right]^{1/2} \tag{2.54}$$

This gives us an expression for x_n in terms of the applied voltage, which can be used to calculate C_j. Taking the derivative with respect to applied voltage and rearranging terms gives

$$\frac{dx_\mathrm{n}}{dV_\mathrm{a}} = \frac{1}{N_\mathrm{d}}\left[\frac{\varepsilon_\mathrm{s}}{2q(1/N_\mathrm{a} + 1/N_\mathrm{d})(V_\mathrm{bi} - V_\mathrm{a})}\right]^{1/2} \tag{2.55}$$

and the junction capacitance is therefore

[44] Equations (2.48) and (2.51) show that the time delays associated with the minority carriers can be thought of in terms of an equivalent *RC* time constant. This implies that a short diode typically responds more quickly than a long diode (since x_B, $x_\mathrm{E} \ll L_p$, L_n).

$$C_j = \left[\frac{q\varepsilon_s}{2(1/N_a + 1/N_d)(V_{bi} - V_a)}\right]^{1/2} = \frac{\varepsilon_s}{x_d} \tag{2.56}$$[45]

Thus for $|V_a|$ much greater than V_{bi}, the capacitance of a *pn* step junction varies approximately inversely with the square root of the reverse bias. Devices which employ this voltage variable capacitance are called *varactor* diodes (variable reactor). Varactors find application in filters, oscillators, tuning circuits, etc. Note that the dependence of capacitance on reverse bias will be determined by the doping profile near the junction (since this will determine the distribution of charge in the space-charge region) and can be tailored for specific applications. Generally, the *pn* junction capacitance can be written

$$C_j = \frac{C_{j0}}{\left[1 - \frac{V_a}{V_{bi}}\right]^m} \tag{2.57}$$

where C_{j0} is the zero-bias junction capacitance and m depends on the doping profile of the junction. We have seen that $m = 1/2$ for a step junction. For a linearly graded junction $m = 1/3$, etc. (see [2] for further details).

2.1.4.4 Total Junction Charge Storage

In general, we can see that the relative importance of C_j and C_d depends strongly on the applied junction voltage: Under reverse bias the junction capacitance dominates, but under forward bias the exponential factor in C_d makes diffusion capacitance important as well. For low-to-moderate forward bias it is necessary to consider both types of charge storage in order to obtain an accurate value for the total *pn* junction capacitance as illustrated in Fig. 2.15. For applications where there is a low-voltage sinusoidal excitation of the diode that is biased at a dc operating point, the above analysis thus results in the *pn* diode small-signal equivalent circuit shown in Fig. 2.16.

It is important to discuss some of the limitations of the small-signal parameters presented above: First of all, it is clear that the exponential dependence of the diffusion capacitance cannot be expected to grow indefinitely as the applied bias approaches the built-in potential, just as the ideal diode current (and hence conductance) did not (see Fig. 2.14). In addition, at large forward bias the depletion layer is effectively eliminated and thus the basis for junction capacitance derived above is

[45] This equation has the same form as the capacitance of a parallel plate capacitor with the plates separated by the depletion width. It can be shown that this correspondence holds for arbitrary dopant profiles.

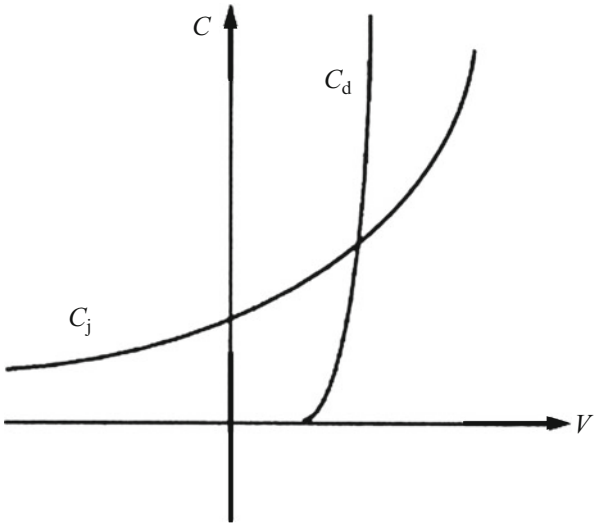


Fig. 2.15 *pn* junction capacitance vs. voltage showing contributions of space-charge (junction capacitance) and excess minority carriers (diffusion capacitance)

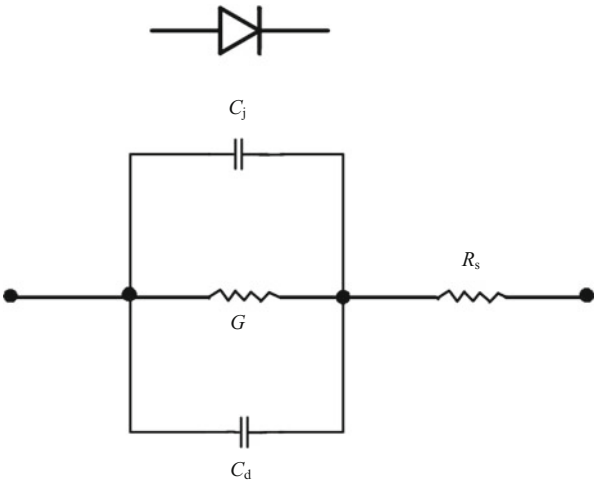


Fig. 2.16 *pn* junction small-signal equivalent circuit. A parasitic series resistance R_s has also been included

no longer valid. For large currents, much of the applied bias will be dropped in the neutral regions of the *pn* junction and this ultimately limits the total junction capacitance at large biases as shown in Fig. 2.17.

A further limitation arises in relation to the frequency of the small-signal excitation applied to the junction. As frequency is increased, a parasitic inductance must also be added to the small-signal equivalent circuit of the diode. The inductive element becomes particularly important at larger forward biases when the capacitive and resistive components of the junction impedance decrease. In addition, the equivalent circuit of Fig. 2.16 should be considered valid only for excitation

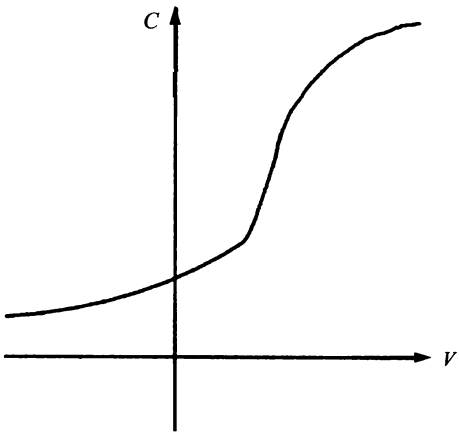


Fig. 2.17 Total *pn* junction capacitance vs. applied bias showing high voltage roll-off

frequencies whose period of oscillation is large compared to the minority carrier lifetime or transit time under forward bias for long- or short-base diodes, respectively. Beyond such frequencies[46] lumped element equivalent circuit models may no longer be applicable and one should employ more detailed distributed calculations that model the entire device by breaking it up into small segments to define a computational grid.[47]

2.1.5 *Transient Behavior and (Large Signal) Diode Switching*

Switching of a *pn* junction from forward bias (a large current or "on" state) to reverse bias (a very small current, or "off" state) and vice versa is a very common process in applications involving diodes. We can gain insight into the on/off switching behavior of a *pn* junction by considering the buildup and decay of Q_p (the hole charge in the n-region).

Consider the distribution of holes in the n-region of an initially unbiased long-base *pn* diode to which a positive constant-current source is suddenly applied: Current can start to flow across the junction very quickly, but before the steady-state

[46] It is possible to define a *diode cutoff frequency* as $f_T = (2\pi\, RC)^{-1}$, where R and C denote the overall (including parasitics, except for inductance) resistance and capacitance components that are dominant at a particular bias, respectively. f_T can be considered an upper limit to how quickly the diode can respond to small-signal excitations.

[47] Ultimately, this can lead to so-called first-principles calculations that consider the individual atoms making up the device. Such calculations typically require very large computational times, but fortunately this kind of precision is not usually required for most devices at present.

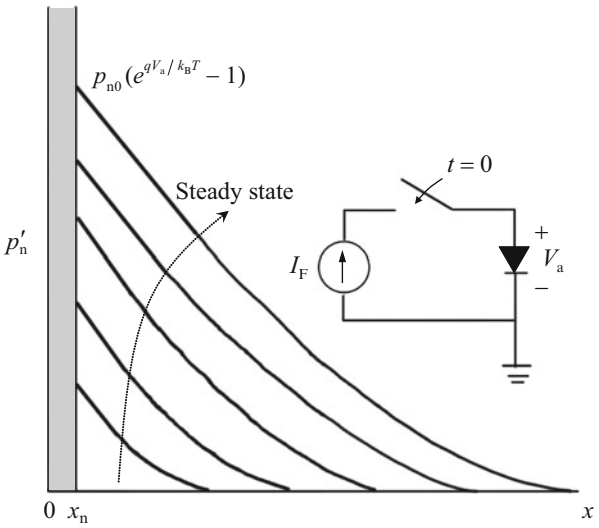


Fig. 2.18 *pn* junction turn-on transient. The slope at the edge of the space-charge region remains constant as the charge increases towards its steady-state distribution (which depends on the current flow and voltage across the junction according to the ideal diode analysis). Current begins to flow very quickly after the current source is connected since only a small number of carriers need to be injected. The long-base diode case is shown (the short-base diode is similar but with a linear triangular distribution instead of the decaying exponential)

charge distribution can be reached holes must be transported into the neutral n-region. The stored charge Q_p increases with time as holes are supplied and the voltage across the junction increases toward its steady-state value. A sketch of the transient increase of stored holes in this situation would have the form shown in Fig. 2.18.

The time required to reach steady state is given by the ratio of the steady-state stored hole charge to the size of the current source (neglecting electrons in the p-side[48]):

$$Q_p(\infty)/I_F = \tau_t \,(\text{short-base}) \Rightarrow \tau_p/2 \,(\text{long-base}) \qquad (2.58)^{49}$$

On the other hand, the turnoff time of the diode is limited by the speed at which stored carriers can be removed from the neutral regions: When a reverse bias is

[48] In other words, we are considering a p^+n diode. Non-idealities (recombination in the space-charge region, etc.) are also ignored in this treatment.

[49] This result can be obtained by noting that both the stored charge and the current flowing through the junction can be expressed in terms of the bias appearing across the junction (via exponential factors) using the results of the ideal diode analysis carried out earlier.

suddenly applied across the forward-biased junction the current can reverse direction quickly because the gradient near the edge of the space-charge region can change with only a small change in the number of stored carriers as illustrated in Fig. 2.19a.

The dc forward current just before switching ($t = 0$–) is given by

$$I_{\mathrm{F}} = \frac{V_F - V_{\mathrm{a}}}{R_F} \approx \frac{V_F}{R_F} \tag{2.59}$$

where the last expression assumes the voltage source is much greater than the voltage appearing across the diode. Similarly, the magnitude of the reverse current immediately after switching ($t = 0$+) is given by

$$I_{\mathrm{R}} = \frac{V_R + V_{\mathrm{a}}(t)}{R_R} \approx \frac{V_R}{R_R} \tag{2.60}$$

once again assuming a large source voltage in the latter expression.

Thus the diode is able to conduct a large amount of current in the reverse direction and the junction will remain forward biased until the injected minority carrier charge near the edge of the depletion region is removed. A plot of current versus time is initially nearly constant until the excess minority carrier concentration at the edge of the depletion layer gets close to zero (Fig. 2.19b), which is termed the diode *storage time*, t_{s}.

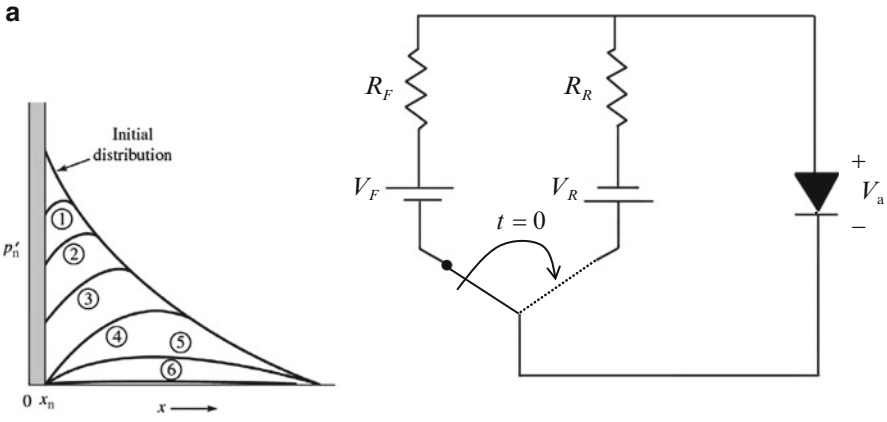


Fig. 2.19 (**a**) *pn* junction behavior during switch off. The initial forward-bias steady minority hole carrier distribution (sketched here for a long-base diode) decreases vs. time; however, the junction can still conduct current in the reverse direction immediately after switching since the slope at the edge of the depletion region can change direction. After a certain storage time, t_{s}, the current begins to decay strongly. (After [4]) (**b**) Current vs. time during switching off of a *pn* junction, illustrating storage time, t_{s}, and recovery time t_{r}, as the voltage across the junction changes from positive to negative. The corresponding decay of the excess minority carrier distributions for a long-base diode is also shown (After [7])

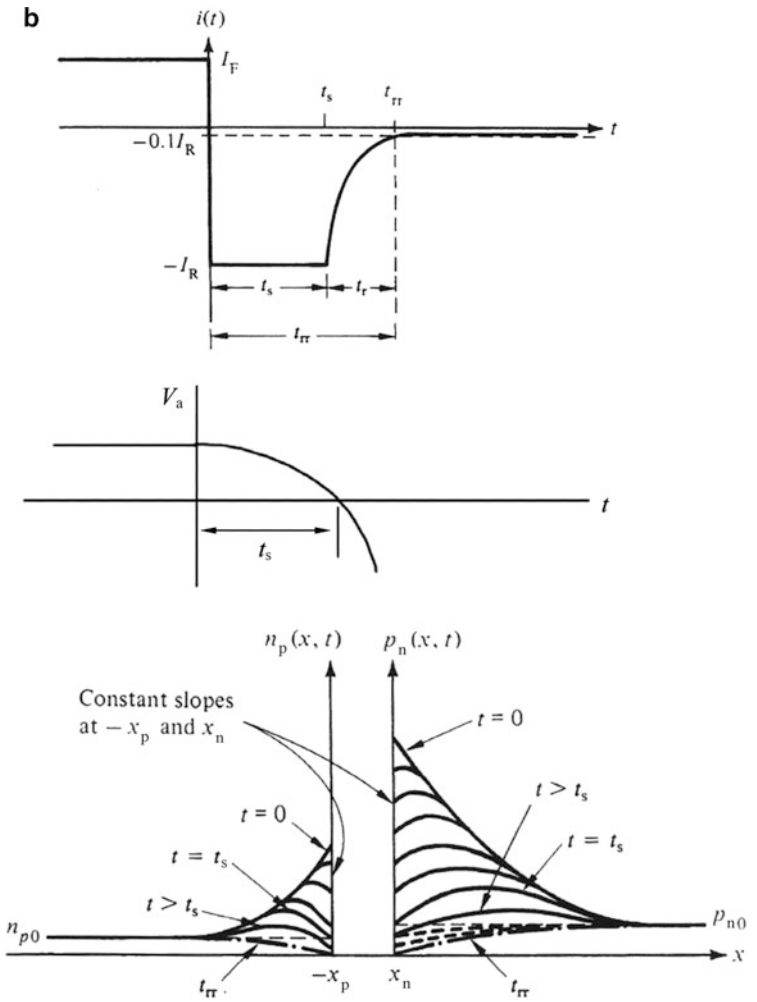


Fig. 2.19 (continued)

The current then decays on a timescale (denoted by the *recovery time*,[50] t_r) mainly determined by carrier lifetime in the neutral region.

A rough estimate of the storage time for a long-base diode can be obtained by noting

$$t_s \approx \frac{\delta \cdot Q_p}{I_R} = \frac{\tau_p}{2} \frac{I_F}{I_R} \cdot \delta \tag{2.61}$$

[50] The recovery time is usually taken as the point at which the reverse current has dropped to 10 % of its initial value.

where Eq. (2.58) has been employed and here δ represents the fraction of the initial charge Q_p that is removed before the reverse current begins to decay.[51] Accurate expressions for the storage and recovery times can be determined by solving the time-dependent continuity equation [3]. For a long-base p^+n junction the storage time turns out to be determined by

$$\text{erf}\sqrt{\frac{t_s}{\tau_p}} = \frac{I_F}{I_F + I_R} \tag{2.62a}$$

where erf (x) is the error function. An approximate analytical solution can also be found as

$$t_s \approx \tau_p \ln\left[1 + \frac{I_F}{I_R}\right] \tag{2.62b}$$[52]

The continuity equation solution for the recovery time t_r, results in

$$\text{erf}\sqrt{\frac{t_r}{\tau_p}} + \frac{\exp\left(-t_r/\tau_p\right)}{\sqrt{\pi t_r/\tau_p}} = 1 + 0.1\left(\frac{I_R}{I_F}\right) \tag{2.62c}$$

In summary, even without any complex mathematical models we can see that in order to switch a *pn* diode quickly we need to be able to produce a large reverse current as well as have a small minority carrier lifetime in the neutral region. In general, a fast *pn* diode should minimize charge storage under forward bias. Another way to accomplish this is by employing a short-base diode whose switching response will instead be *transit time limited*.[53]

Note that we neglected the majority carriers in the above discussion because they respond to changes in electric fields much more quickly than minority carriers since they do not need to recombine. Majority carriers respond within the so-called *dielectric relaxation time*, which is usually on the order of less than a picosecond in silicon. This is also the reason that the small-signal junction capacitance derived earlier is less sensitive to frequency than the forward-bias diffusion capacitance.

[51] Although an analytical expression for this parameter based on the diode material properties is possible, empirically a value of $\delta \sim 0.2$ agrees quite well with data over a fairly broad range of currents.

[52] This equation overestimates the storage time by a progressively larger amount as I_R/I_F increases.

[53] Equations (2.59), (2.60), (2.61), (2.62a), (2.62b), and (2.62c) also apply to a short-base diode if τ_p is replaced by $2\tau_t$.

2.1.5.1 Equivalent *pn* Diode Circuits for Transient Problems

The nature of charge storage at the *pn* junction complicates the use of equivalent circuits for hand calculations involving transient problems. For example, switching diodes that are sequentially forward and reverse biased during circuit operation will have the dominant contribution to charge storage shift between diffusion and depletion layer charges during the transient itself, as we saw above.

If an accurate picture of the current and voltage as functions of time is required the diode can be *piecewise-linearly* approximated, i.e., the nonlinear charge storage and conductance effects can be approximated to first order by linear elements over a small voltage range. If the voltage increment is made small enough, this approximation is accurate and arbitrary precision can be obtained by joining together a sufficient number of piecewise-linear approximations to represent the entire voltage variation in a given transient problem, typically using a computer. The small-signal parameters we found earlier are the piecewise-linear approximations we need in order to represent charge storage and conductance in the *pn* junction for this type of calculation.

Panel 2.2: *pn* Junction Fabrication and Integrated Circuits To form a *pn* junction diode using planar silicon technology (see Appendix A, (sect. A.3)) it is only necessary to diffuse a p-type region into an n-type wafer and make electrical contact to the front and back of the wafer as shown in Fig. 2.20a. However, in a modern integrated circuit (IC) there will in general be many diodes and other devices on the same wafer and we must be able to electrically isolate these devices from each other for proper circuit operation. This is accomplished by using a combination of insulating oxide regions and *pn* junctions that are kept under reverse bias at all times.[54] For example, to form an isolated array of diodes the sequence of steps shown in Fig. 2.20b can be performed.

One problem with IC diodes and other devices formed in a similar manner is so-called *current crowding*, which reduces the effective area of the diode and limits its current carrying capacity (Fig. 2.21a). To get around this difficulty and allow current to flow vertically through the diode junction in a uniform manner, a heavily doped, low-resistance *buried layer* is often diffused onto the substrate. The heavily doped n-type (or p-type) region is formed before growing the epitaxial[55] layer and can be included beneath each junction region as illustrated in Fig. 2.21b whenever vertical currents (i.e., perpendicular to the substrate) are important to integrated

[54] These critical IC interconnection techniques were first developed by Noyce and Lehovec working independently in 1959.

[55] Epitaxy is a type of thin film crystal growth. See Appendix A, Sect. A.3, for further details.

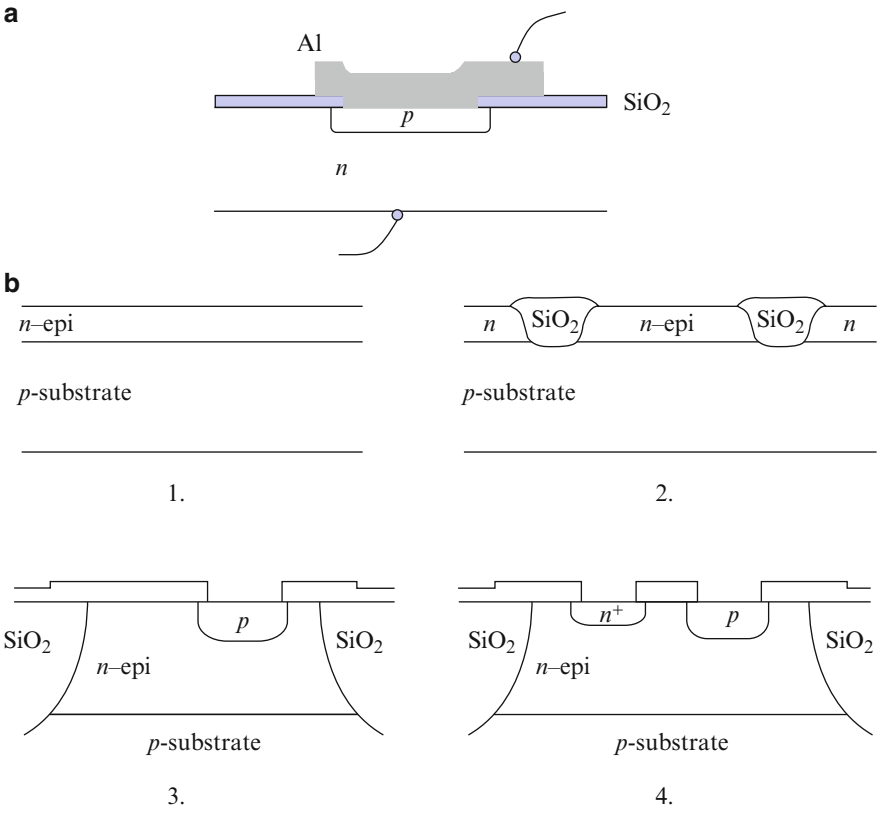


Fig. 2.20 (**a**) Typical planar *pn* junction discrete device structure. (**b**) IC isolated diode array fabrication sequence using an n-type epitaxial layer grown on a p-type substrate. The n^+ layer in image 4 is for making a low-resistance contact to the epitaxial layer as described in Sect. 2.2 and 2.3. Note that the oxide formed above the epitaxial layer is thicker in order to prevent the unintentional formation of field-induced conducting surface channels (see Chap. 4) and is referred to as a *field oxide* (Adapted from [4])

circuit device operation. A modern IC *pn* junction (Fig. 2.21c) employs these and many other advances to create the very precisely controlled structures required for state-of-the-art electronics.

2.2 Metal–Semiconductor Junctions

Most electronic devices are interconnected using metallic wiring that forms metal–semiconductor contacts. For example in an integrated circuit there are typically many millions of metal contacts to silicon. The properties of these contacts can

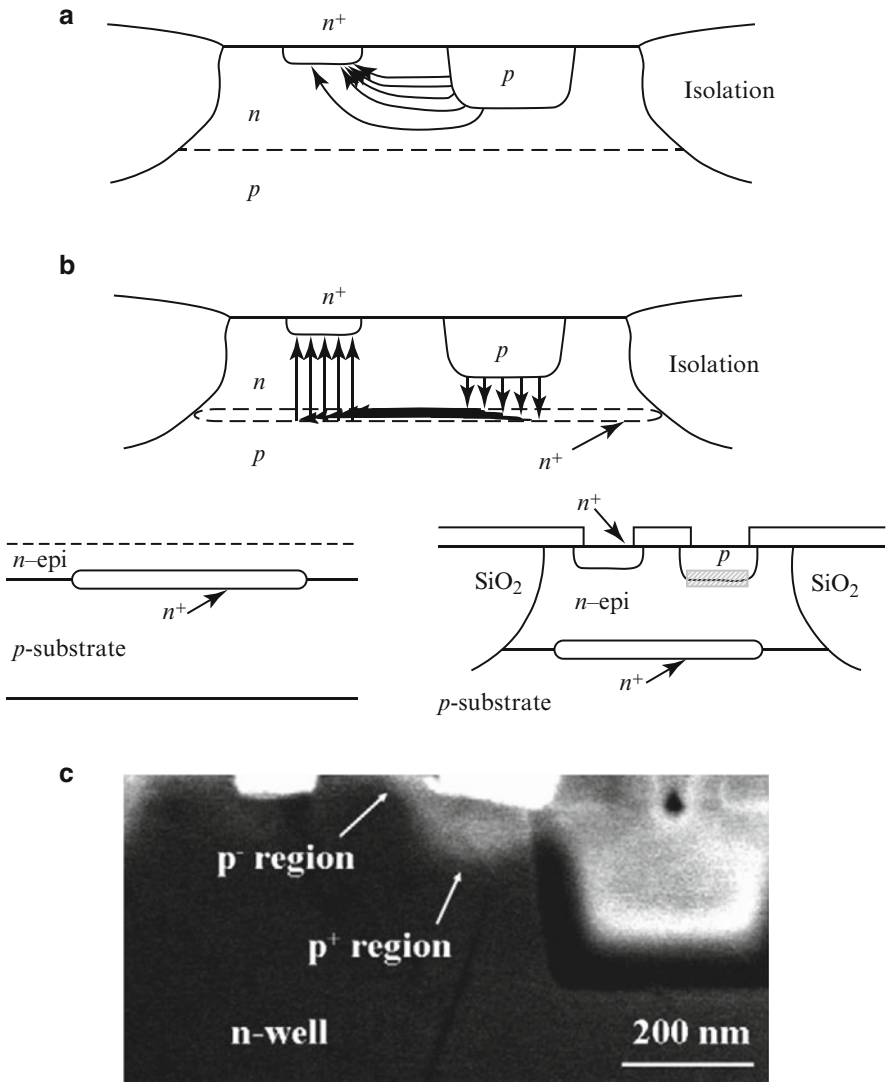


Fig. 2.21 (**a**) IC *pn* junction current crowding phenomenon. (**b**) IC *pn* junction heavily doped buried layer fabrication process to prevent current crowding. The active *pn* junction device interface is highlighted by hatched region (Adapted from [4]). (**c**) Cross section of a typical *pn* junction structure used in integrated circuits (After J.-H. Lee, P.-T. Liu, Microelectronic Engineering **95**, 5 (2012))

vary considerably and we need to understand the nature of thermal equilibrium that is established when a metal–semiconductor junction is formed. The concepts and analysis of metal–semiconductor junctions are similar to what we used to study *pn* junctions and often only need minor modification.

The first metal–semiconductor junctions studied were in the form of point-contact structures: In 1874 Braun was able to realize both rectifying and low-resistance contacts and around the same time Schuster observed rectification in a contact made using similar techniques. The basic theory of metal–semiconductor junctions was developed around 1938 by Schottky along with related work by Mott.[56]

2.2.1 Metal–Semiconductor Barriers (Blocking Contacts)

2.2.1.1 Thermal Equilibrium

We once again use the fact that the Fermi level must be constant in thermal equilibrium to construct the band edge diagram. In the case of a copper contact to n-type silicon, before the two materials are brought together we have the situation depicted in Fig. 2.22a. The relative values of the work functions (and hence Fermi levels) of the isolated materials determine the band edge diagram of the metal–semiconductor junction in thermal equilibrium just as they did for the *pn* junction. For the case shown in the figure electrons from the semiconductor (higher Fermi level) will be transferred into the metal (lower Fermi level) in order to achieve equilibrium. Figure 2.22b shows the resulting thermal equilibrium band edge diagram from which we see that the height of the barrier step[57] at the interface is given by

$$q\phi_B = q(\Phi_M - \chi) \tag{2.63a}$$[58]

where the built-in potential of the semiconductor is found from

$$\phi_{bi} = \Phi_M - \Phi_S \tag{2.63b}$$

The transfer of electrons from the semiconductor into the metal results in a built-in electric field at the junction due to the uncompensated donor ions near the interface, which creates a space-charge region as shown in Fig. 2.23a (cf. Fig. 2.3b for a *pn* junction). We are once again ignoring any contribution to the charge density from free electrons and holes in the semiconductor, i.e., the depletion approximation is assumed to hold. In the metal there exists a thin layer of negative interfacial or surface charge[59] that is equal in magnitude to the space charge in the semiconductor.

[56] Many of the important features of metal–semiconductor junctions (and *pn* junctions) were also developed theoretically by Davydov between 1938 and 1939.

[57] Known as the Schottky barrier.

[58] The analogous barrier height for a p-type semiconductor is given by $q\phi_B = E_g - q(\Phi_M - \chi)$.

[59] Recall from electrostatics that free extra charge cannot exist in the interior of a (metallic) conductor.

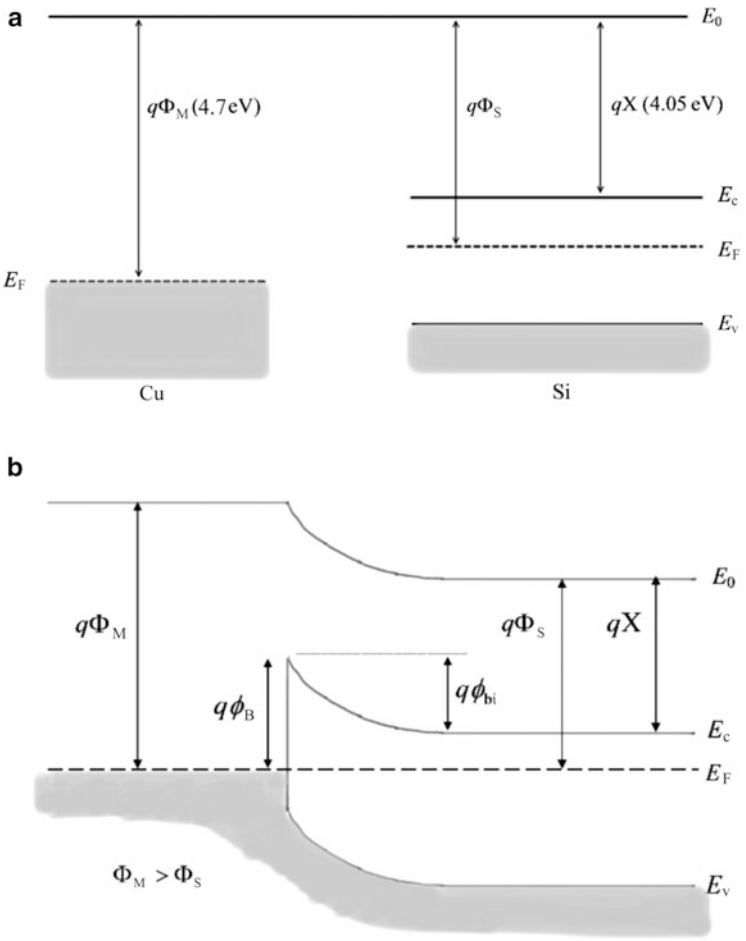


Fig. 2.22 (**a**) Copper and n-type silicon band edge diagrams in isolation. (**b**) Copper–silicon junction thermal equilibrium band edge diagram. Note the constancy of the electron affinity in the semiconductor regardless of position. In this example electrons were transferred from the semiconductor and therefore the bands bend upward at the interface. For the corresponding junction to a p-type semiconductor electrons are instead transferred into the semiconductor and the bands bend downward

We can now analyze the metal–semiconductor interface using electrostatics as we did for the *pn* junction. The electric field is thus given by

$$E_x = -\frac{qN_d}{\varepsilon_s}(x_d - x), \quad 0 \le x \le x_d \tag{2.64}$$

and plotted in Fig. 2.23b. The maximum field, located at the metallurgical junction, has the value

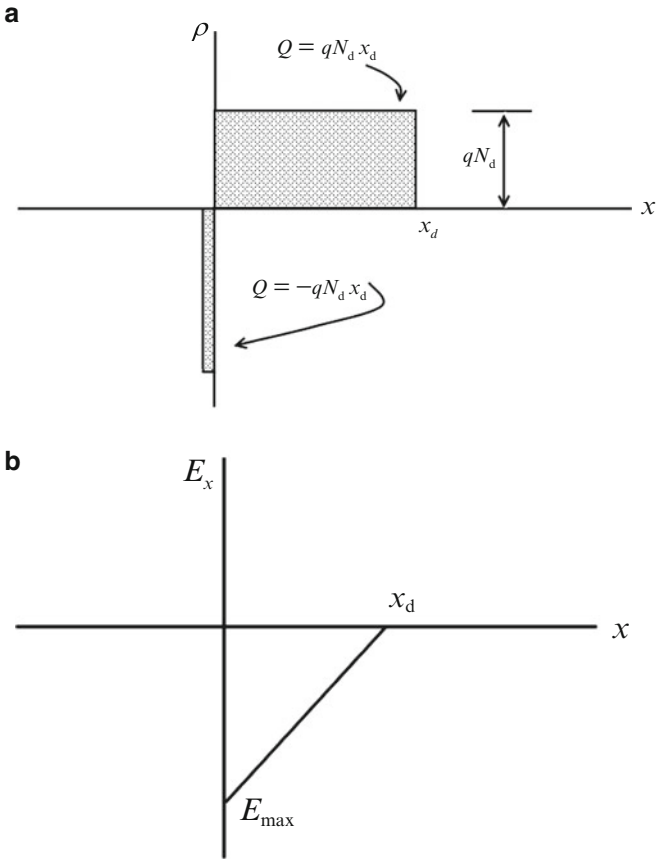


Fig. 2.23 (**a**) Space-charge density for metal–semiconductor junction in Fig. 2.22 (charge per unit area Q is also shown). (**b**) Corresponding electric field in depletion region of metal–semiconductor junction

$$E_{\max} = \frac{-qN_{\mathrm{d}}x_{\mathrm{d}}}{\varepsilon_{\mathrm{s}}} \tag{2.65}$$

The potential variation in the semiconductor is given by

$$V(x) = \phi_{\mathrm{bi}} - \frac{qN_{\mathrm{d}}}{2\varepsilon_{\mathrm{s}}}(x_{\mathrm{d}} - x)^2, \quad 0 \le x \le x_{\mathrm{d}} \tag{2.66}$$

where the built-in potential, expressed as the negative of the area under the field curve, is

$$\phi_{\mathrm{bi}} = -\frac{1}{2}E_{\max}x_{\mathrm{d}} = \frac{1}{2}\frac{qN_{\mathrm{d}}x_{\mathrm{d}}^2}{\varepsilon_{\mathrm{s}}} \tag{2.67}$$

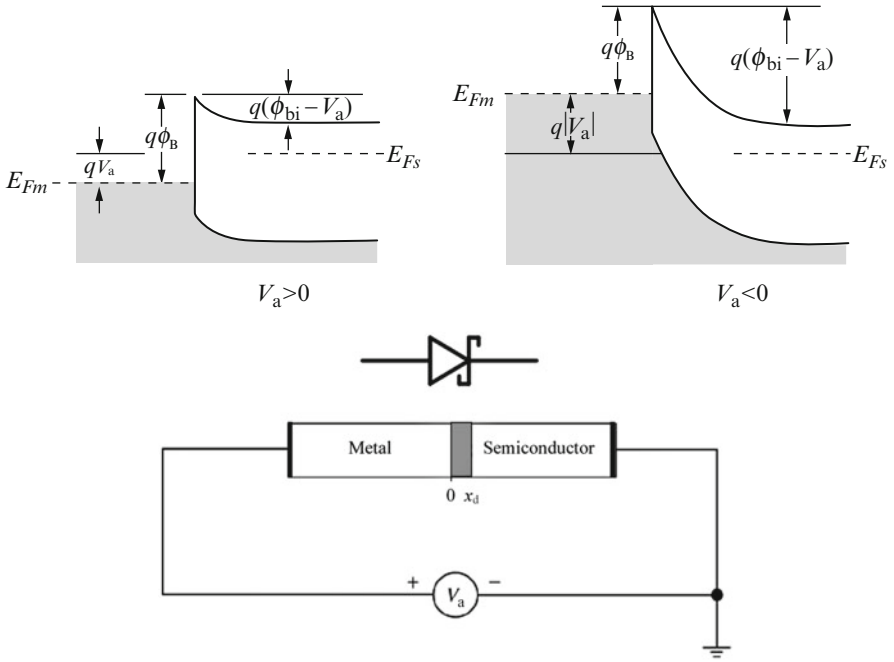


Fig. 2.24 Effect of applied bias on metal–semiconductor (n-type) junction potential barrier in forward and reverse directions. The interface barrier height ϕ_B is assumed unaffected. Low-resistance contacts are made to both the metal and semiconductor regions and the applied bias is assumed to drop entirely across the semiconductor depletion layer (For a Schottky barrier to a p-type semiconductor, the forward- and reverse-bias polarities are interchanged) (Electrical symbol also shown)

Using this expression we can write the depletion width as

$$x_d = \sqrt{2\phi_{bi}\epsilon_s/qN_d} \tag{2.68}$$

Note that the results derived above in Eqs. (2.64), (2.65), (2.66), (2.67), and (2.68) are identical to those for a p^+n junction. Similarly, a metal–semiconductor barrier based on p-type material would be analogous to a n^+p junction.

2.2.1.2 *I–V* Characteristic

The potential barrier at the interface, ϕ_B, makes it difficult for free electrons to travel from the metal to the semiconductor. To first order this barrier is independent of applied bias since it is determined by the material properties of the metal/semiconductor and not any built-in charges. On the other hand, the built-in potential, ϕ_{bi}, impeding electrons from traveling from the semiconductor to the metal can be altered from its thermal equilibrium value by an applied bias. Analogous to the *pn* junction, the built-in potential is reduced when the metal is biased positively with respect to the semiconductor, and it is increased when the metal is made more negative as illustrated in Fig. 2.24.

The applied bias changes the depletion width to

$$x_{\mathrm{d}} = \sqrt{2(\phi_{\mathrm{bi}} - V_{\mathrm{a}})\varepsilon_{\mathrm{s}}/qN_{\mathrm{d}}} \tag{2.68}$$

Similarly, the magnitude of the maximum electric field in the junction becomes

$$E_{\max} = \frac{2(\phi_{\mathrm{bi}} - V_{\mathrm{a}})}{x_{\mathrm{d}}} \tag{2.69}$$

which is equivalent to Eq. (2.32b).

To obtain an expression for the metal–semiconductor barrier *I–V* characteristic we make use of the fact that in thermal equilibrium the net current across the metal–semiconductor interface must be zero. In other words, the rate at which electrons cross over the barrier into the semiconductor from the metal is balanced by the rate at which electrons cross the barrier into the metal from the semiconductor. We can apply these arguments at the boundary plane of the band edge diagram in thermal equilibrium:

The electron currents in either direction across the junction boundary due to thermal motion of carriers are proportional to the density of electrons at the boundary.[60] In the semiconductor this density is (referring to Fig. 2.22b)

$$n_{\mathrm{s}} = N_{\mathrm{c}}\exp\left(-\frac{q\phi_{\mathrm{B}}}{k_{\mathrm{B}}T}\right) = N_{\mathrm{d}}\exp\left(-\frac{q\phi_{\mathrm{bi}}}{k_{\mathrm{B}}T}\right) \tag{2.70}$$

Thus the condition for thermal equilibrium at the junction corresponds to

$$|J_{\mathrm{M\text{-}S}}| = |J_{\mathrm{S\text{-}M}}| = KN_{\mathrm{d}}\exp\left(-\frac{q\phi_{\mathrm{bi}}}{k_{\mathrm{B}}T}\right) \tag{2.71}$$

where $J_{\mathrm{M\text{-}S}}$ and $J_{\mathrm{S\text{-}M}}$ are the thermally induced current densities directed from the metal toward the semiconductor and vice versa, and K is a constant of proportionality. When a bias V_{a} is applied the built-in potential of the semiconductor is changed and we can expect the flux of electrons from the semiconductor toward the metal to be modified by a factor

$$n_{\mathrm{s}} = N_{\mathrm{d}}\exp\left(-\frac{q(\phi_{\mathrm{bi}} - V_{\mathrm{a}})}{k_{\mathrm{B}}T}\right) \tag{2.72}$$

The flux of electrons from the metal to the semiconductor, however, is not affected by the applied bias because the barrier at the interface is assumed to remain fixed at its equilibrium value. We can subtract these two components to

[60] For a system obeying Maxwell–Boltzmann statistics, it can be shown that the electron (or hole) current across a plane due to carriers in thermal motion can be expressed as $J = qn\overline{v}_{\mathrm{th}}/4$, where $\overline{v}_{\mathrm{th}}$ is the mean or average thermal velocity, $\overline{v}_{\mathrm{th}} = \sqrt{\frac{8k_{\mathrm{B}}T}{\pi m^*}}$.

obtain an expression for the net current density from the metal into the semiconductor under applied bias:

$$\begin{aligned} J &= J_{\text{M-S}} - J_{\text{S-M}} \\ &= KN_{\text{d}}\exp\left(-\frac{q(\phi_{\text{bi}} - V_{\text{a}})}{k_{\text{B}}T}\right) - KN_{\text{d}}\exp\left(-\frac{q\phi_{\text{bi}}}{k_{\text{B}}T}\right) \end{aligned} \tag{2.73}$$

which can be written

$$J = J_0[\exp(qV_{\text{a}}/k_{\text{B}}T) - 1] \tag{2.74a}$$

where $J_0 = KN_{\text{d}}\exp(-q\phi_{\text{bi}}/k_{\text{B}}T)$. Equation (2.74a) is in the form of the ideal diode equation, once again showing the strong asymmetry or rectification of current flow across the junction. Thus, this type of metal–semiconductor junction is referred to as a *Schottky diode*.

Given the resemblance to a *pn* junction it is not surprising that the Schottky diode has a similar current–voltage dependence although the presence of the metal–semiconductor barrier causes the expression for saturation current density to be quite different. If we repeat the above analysis starting from Eq. (2.70) but instead work with the expression involving N_{c} and rewrite Eq. (2.71) in terms of $J = qn\overline{v}_{\text{th}}/4$, the saturation current density can be found explicitly as

$$J_0 = \frac{4\pi q m_n k_{\text{B}}^2 T^2}{h^3}\exp\left(-\frac{q\phi_B}{k_{\text{B}}T}\right) \tag{2.74b}$$

The model leading to Eqs. (2.74) captures the essential physics of transport across a metal–semiconductor barrier.

The principles of electron transport over the barrier and into the metal were first derived by Bethe in 1942.[61] If scattering inside the depletion region is significant this model needs to be modified to include drift–diffusion processes. Additional transport processes that can occur are tunneling through the barrier (discussed further below), generation–recombination in the depletion region (similar to a *pn* junction), and minority carrier contributions. The relative importance of these effects will depend on the particular metal/semiconductor combination as well as the applied bias. In addition, if the doping profile in the semiconductor is not constant[62] the saturation current density will also be modified. In more detailed treatments the saturation current density is also not completely independent of applied voltage. However, the dependence is weak compared to the exponential term and, as we saw for the *pn* diode, the current–voltage relationship for a metal–semiconductor diode can be approximated quite accurately using Eq. (2.42):

[61] This is known as the thermionic emission model.

[62] For example, in a *Mott barrier* a thin, lightly doped semiconductor region contacts the metal, which transitions to a highly doped bulk region a short distance from the junction.

$$I = I'_0 \left[\exp\left(\frac{qV_a}{\eta k_B T} \right) - 1 \right]$$

where I'_0 is independent of voltage and η is usually found experimentally to be slightly greater than 1 for a Schottky diode.[63]

2.2.2 Metal–Semiconductor Ohmic Contacts (Non-blocking)

In the metal–semiconductor barrier junction examined in the previous section, the applied voltage was dropped mainly across the high-resistance junction region and currents are therefore limited by the metal–semiconductor contact. The opposite case, in which the contact offers negligible resistance to current flow, defines an *ohmic contact*. Being able to make low-resistance metal–semiconductor junctions is very important for electrically contacting devices without modifying their inherent *I–V* characteristics. This is even more important in modern integrated circuit devices which are smaller and subsequently have larger current densities and regions with increased doping levels.

Two standard approaches are used to make ohmic contacts:

1. Tunnel contacts
   The metal–semiconductor junction barrier structure we considered previously, for example, can be made ohmic if the effect of the barrier on current flow is made negligible. This can be achieved by heavily doping the semiconductor so that the barrier width, given by the depletion width in Eq. (2.68), namely,

$$x_d = \sqrt{\frac{2\varepsilon_s \phi_{bi}}{qN_d}}$$

   is made very small. When the barrier width approaches a few nanometers, *tunneling* transport through the barrier can take place when a bias is applied as depicted in Fig. 2.25a. Many electrons are available to take part in tunneling and currents rise very rapidly with applied bias, regardless of polarity. Hence, a metal–semiconductor contact at which tunneling is the significant transport process has a very small resistance and it is virtually always an ohmic contact. To ensure a very thin barrier, the semiconductor is often doped until it is *degenerate* (i.e., until the Fermi level is very close to or enters either the valence or the conduction band). Tunnel contacts to heavily doped semiconductors are widely used in integrated circuits.

[63] At higher doping levels ($\sim 10^{18}$ cm^{-3}) and/or lower temperature η begins to deviate from unity [2] as tunneling becomes more important.

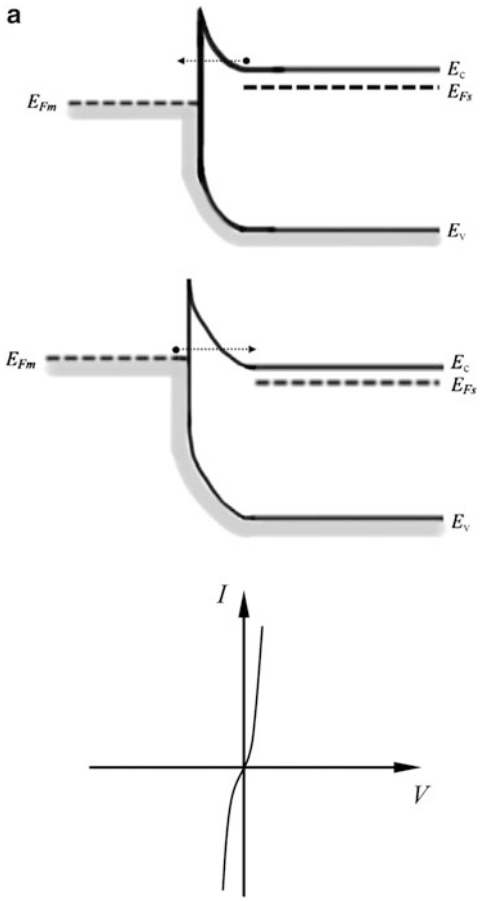


Fig. 2.25 (**a**) Ohmic tunnel contact and corresponding *I–V* characteristic showing rapid current increase with voltage for both polarities due to electron tunneling. (**b**) Schottky ohmic contact. Electrons are transferred from the metal into the semiconductor and therefore the bands bend downward at the interface because of the enhanced electron concentration. For the corresponding ohmic contact to a p-type semiconductor electrons are instead transferred into the metal and the bands bend upward due to the increased hole concentration

2. Schottky ohmic contacts
   Another way to achieve ohmic behavior is to choose materials with the proper work function difference so majority carriers become more numerous near the contact than they are in the bulk of the semiconductor. In our rectifying contact example, the metal had a greater work function than the n-type semiconductor and electrons were therefore transferred from the semiconductor to the metal to reach thermal equilibrium. If we instead choose a metal with a smaller work function, electrons will be transferred from the metal to the semiconductor. In this case the semiconductor surface is not depleted when it comes into equilibrium with the metal, but rather has an enhanced majority carrier concentration

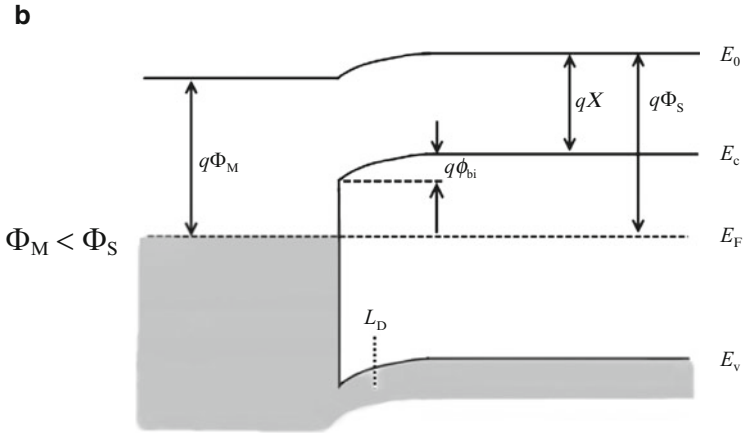


Fig. 2.25 (continued)

(Fig. 2.25b). Note in this diagram the *Debye length*, L_D, is a measure of the spatial extent of the additional charge carriers near the semiconductor interface.[64]

We can summarize Schottky contacts thusly; metal contacts to n-type material can be classified as being *ohmic* when the bands bend *down* toward the interface and rectifying or *blocking* when the bands bend *up*. The inverse conditions apply for contacts to p-type material[65] (see Figs. 2.22b and 2.25b).

The essential condition for ohmic behavior is an unimpeded transfer of charge carriers between the two materials forming the contact: At contacts there are generally built-in potentials and unless very thin barriers are present (i.e., tunneling is significant) majority carriers must be more numerous than they are in the bulk in order to achieve an ohmic contact.

[64] This may also be viewed as a measure of the *screening length* or the distance over which the free charges in the semiconductor rearrange themselves in response to the additional electrons in order to cancel out the electric field inside the bulk of the semiconductor. The Debye length is proportional to the density of free charges in a material; for most metals L_D is typically well below 1 nm, while it can range from 10 to 100 nm or more in a semiconductor depending on the doping level, with a maximum value for the intrinsic case.

[65] Note that the equations given for a metal contact to n-type material apply equally for contacts to p-type material with appropriate substitutions for the doping type (N_a for N_d), effective mass (m_p for m_n), etc.

2.2.3 *Deviations from Ideal Behavior*

2.2.3.1 Surface States and Barrier Height

The metal–semiconductor junctions considered thus far were ideal in the sense that we assumed the allowed energy states at the junction interface surface were the same as those in the bulk semiconductor material. In general, any surface of a crystal will introduce extra allowed states for electrons, often referred to as Shockley–Tamm sates. These states arise because the electrons on the surface are bonded only from the side directed toward the bulk as shown in Fig. 2.26a. In addition, impurities and crystal defects are sources of other types of surface states.

In real metal–semiconductor junctions the surface states almost always modify the height of the junction barrier from the idealized case. To account for these interfacial effects the metal–semiconductor junction can be treated as containing a very thin intermediate region sandwiched between the two materials: In Fig. 2.26b the n-type semiconductor is depleted of electrons near its surface by acceptor surface states. If a rectifying contact is now formed with a metal having a larger work function, electrons will be transferred from the surface states to the metal. However, since the density of surface states is usually very large (proportional to the density of atoms at the interface), a negligible movement of the Fermi level at the semiconductor surface transfers sufficient charge to equalize the Fermi levels. This is referred to as *Fermi-level pinning*.[66]

When the Fermi level is pinned the barrier height of a metal–semiconductor junction becomes

$$q\phi_B = \left(E_g - q\phi_0\right) \tag{2.75}$$

where $q\phi_0 = E_F - E_v$, compared to the ideal case (no surface states) we saw earlier (Eq. 2.63a):

$$q\phi_B = q(\Phi_M - X)$$

In general the barrier height will have an intermediate value that depends on the magnitude of surface states near the Fermi level of the semiconductor. However, in practice, Schottky barrier heights for most standard semiconductors based on the diamond lattice (Si, Ge, GaAs, etc.) are usually more accurately described by Fermi-level pinning with a smaller dependence on metal work function. Experimentally, it is found for these semiconductors,

[66] The importance of surface states on metal–semiconductor interfaces was pointed out by Bardeen in 1947. The effect of Fermi-level pinning is somewhat analogous to adding a very thin heavily doped layer between the metal and semiconductor.

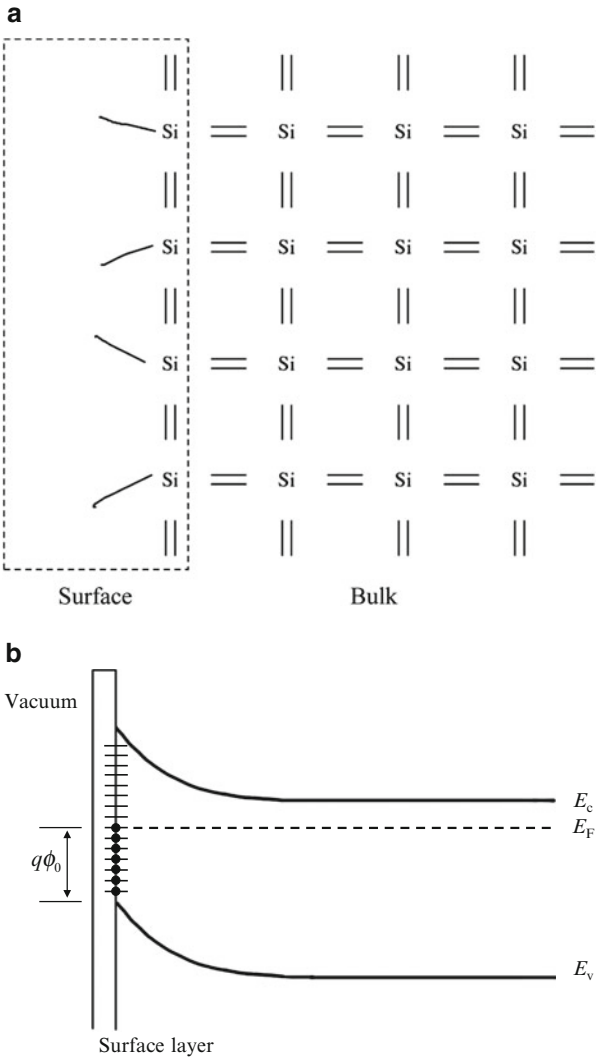


Fig. 2.26 (**a**) Illustration of incomplete or dangling bonds at the surface of a semiconductor that lead to a large density of localized surface interface states. (**b**) Surface transition layer representing semiconductor interface states leading to a space-charge region near the surface of the semiconductor at thermal equilibrium

$$q\phi_0 \approx \frac{1}{3}E_g \tag{2.76}$$

so that the barrier height $q\phi_B$ is roughly 2/3 of the band gap energy. Although surface states and other non-idealities will modify the exact Schottky barrier height, the ideas and properties of idealized metal–semiconductor junctions discussed

above are still valid[67] as long as an accurate barrier height is used, obtained from experimental data or other means.

2.2.3.2 Effects at High and Low Applied Bias

Other deviations from ideal behavior, similar to those observed in *pn* junctions, can also occur in Schottky barrier diodes:

Under large reverse bias avalanche breakdown is generally observed for Schottky diodes based on moderately doped semiconductors, whereas breakdown via tunneling can occur for more heavily doped diodes. Generation currents arising in the depletion region can also cause the reverse saturation current in a Schottky diode to gradually increase with the magnitude of reverse bias. Similarly, space-charge region recombination currents under forward bias can also play a role. The deviation from ideal diode exponential behavior caused by space-charge generation–recombination in Schottky diodes is usually small compared to *pn* junctions and is typically more evident in junctions with large barriers and at low temperatures. In addition, unlike the ideal case, the barrier height itself will also vary with applied reverse bias due to the effects of the applied field and image forces.[68] This so-called *Schottky barrier lowering* is more important for small barrier heights and causes the saturation current to increase gradually with reverse bias. Finally, under large applied forward bias the series resistance of the semiconductor must also be taken into account.

2.2.4 Small-Signal Parameters

2.2.4.1 Conductance

The small signal conductance of an ideal Schottky diode is equivalent to the *pn* diode expression:

$$G = \frac{dI}{dV_a} = \frac{q}{k_B T} I_0 \left(e^{qV_a/k_B T} \right) = \frac{q}{k_B T} (I + I_0)$$

with the only difference being the particular value of the saturation current.

[67] Essentially only Eqs. (2.63a) and (2.63b) will require modification.

[68] Electrons that are emitted from the metal into the semiconductor under reverse bias will induce images charges of the opposite sign in the planar metal surface near the interface, which causes the barrier height to be lowered within a few nanometers from the metallurgical junction.

2.2.4.2 Junction Capacitance

The small-signal junction capacitance per unit area for a Schottky diode can be found using the space charge stored in the semiconductor:

$$Q_s = \sqrt{2q\varepsilon_s N_d(\phi_{bi} - V_a)} \tag{2.77}$$

and therefore,

$$C = \left|\frac{dQ_s}{dV_a}\right| = \sqrt{\frac{q\varepsilon_s N_d}{2(\phi_{bi} - V_a)}} = \frac{\varepsilon_s}{x_d} \tag{2.78}$$

Solving this equation for the total voltage across the junction gives

$$(\phi_{bi} - V_a) = \frac{q\varepsilon_s N_d}{2C^2} \tag{2.79}$$

This indicates that a plot of $1/C^2$ versus the applied voltage should be a straight line. The slope can be used to obtain the doping level in the semiconductor and the intercept with the voltage axis should equal the built-in voltage.[69] Such plots are often used to study semiconductors. In practice, the built-in voltage obtained is not as accurate as the doping level. Commercial profilers can also use this data to plot dopant concentration as a function of position on a semiconductor wafer.

The small-signal equivalent circuit for the Schottky barrier diode is identical to the *pn* diode circuit shown in Fig. 2.16, except for one important difference: Since the Schottky diode is predominantly a majority carrier device, diffusion capacitance due to minority carrier charge storage is absent. This makes the Schottky diode an intrinsically fast device that can respond to frequencies into the THz regime.[70] Similar comments apply to the transient behavior of Schottky diodes: their characteristic timescale is no longer associated with recombination/diffusion processes but rather the transit time associated with the drift of majority carriers in the depletion region, i.e., the dielectric relaxation time. As mentioned earlier, this is a fast process and thus Schottky diodes can be switched very rapidly.[71]

[69] Equation (2.79) is also valid for a one-sided abrupt *pn* junction (p^+n or pn^+).

[70] As for the *pn* diode, we can define a Schottky diode cutoff frequency as $f_T = (2\pi\ RC)^{-1}$. Typically, the fastest Schottky diodes are made from semiconductors with the highest carrier mobility in order to reduce the effect of series resistance.

[71] The external circuit parameters will largely determine how quickly a Schottky diode can be switched as opposed to intrinsic delays in the device itself.

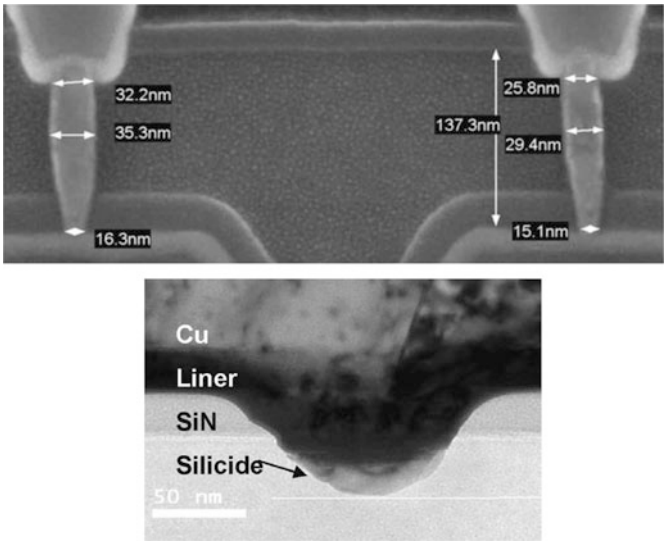


Fig. 2.27 Electron microscope images of IC metal–semiconductor contacts based on copper/silicide conductors (*Source*: K. Ohuchi et al., 8th International workshop on Junction Technology, IWJT, 2008; S.-C. Seo et al., Copper Interconnect Technology Conference, IITC, 2009)

Panel 2.3: Metal–Semiconductor Contact and Device Structures Modern metal–semiconductor contacts and diodes are usually made via planar processing. For creating ohmic tunnel contacts in integrated circuits, a chemical reaction between the metal and underlying silicon contact area is commonly used to form a low-resistance metallic *silicide* (a silicon–metal compound) region, which reduces the lateral resistance of the contact. This process creates a stable, low-resistance electrical connection. Some examples of IC contacts are shown in Fig. 2.27.

Schottky barrier diodes are used extensively in integrated circuits because they are relatively easy to fabricate and combine with other devices on a chip. A typical planar metal–semiconductor diode is shown in Fig. 2.28a. One issue that must often be dealt with in such structures is premature breakdown under moderate levels of reverse bias due to the larger electric fields that exist at the edge of the device.[72] The concentrated electric fields at the edges of the junction can be mitigated in several ways: Fig. 2.28b, c show two common techniques—using an overlapping electrode metal contact or an additional insulating oxide layer to create larger separation at the edges in order to reduce the field. The applications of such structures will be expounded in Sect. 2.4 and discussed further in Chap. 3.

[72] This edge or surface breakdown mechanism can also be important in *pn* diodes but is generally not as severe unless high power devices are required.

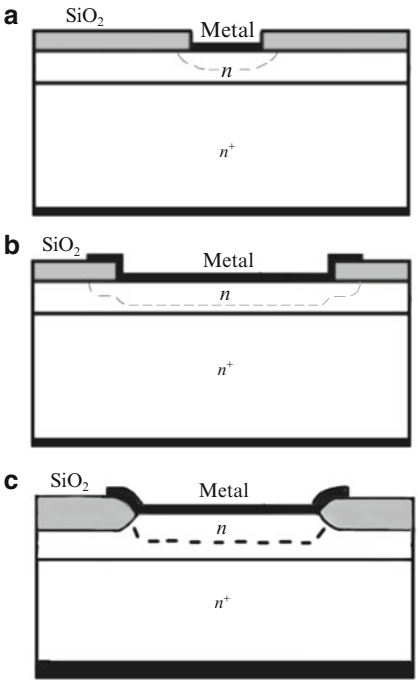


Fig. 2.28 (**a**) Planar Schottky diode device schematic. (**b**, **c**) Approaches for reducing the electric field at the diode edges (*Dotted lines* indicate edge of depletion region) (Adapted from [2])

2.3 Other Types of Junctions

In this section we briefly discuss the properties of two other important junctions that are encountered in practice. Although full details are not provided here, the *pn* and metal–semiconductor results provide sufficient background to understand most other junction types as well.

2.3.1 Isotype Junctions

Figure 2.29 shows the thermal equilibrium band edge diagram for a semiconductor that has two regions with different doping levels that are of the *same type*. Typical examples of these so-called isotype[73] junctions are p^+p and n^+n interfaces. In such junctions, there will be a transfer of carriers (electrons or holes) from the more heavily doped region into the more lightly doped region to achieve thermal equilibrium. The built-in potential of the isotype junction can be found by applying Eq. (2.12).

[73] J. B. Gunn, J. Electron. Control **4**, 17 (1958) is an early study on isotype junctions.

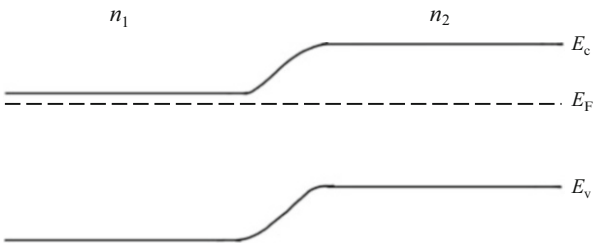


Fig. 2.29 Isotype junction band edge diagram for two n-type regions, $n_1 > n_2$

In contrast to *pn* junctions, however, since the carriers are transferred into a region of the same doping type, there will be an *enhancement* of the carrier concentration near the interface in the more lightly doped side of the junction. This situation is very similar to the ohmic Schottky contact shown in Fig. 2.25b, and indeed an isotype junction may be considered as an ideal ohmic contact in the sense that it consists of the same material (i.e., interface states will not play a role).

2.3.2 *Heterojunctions*

Thus far when considering the junction between two semiconductor regions we have always assumed that it consisted of the same type of semiconductor, known as a *homojunction*. If we lift this restriction, the resulting interface is now called a *heterojunction*. Such junctions began being studied theoretically in the 1950s followed by important experimental studies in the 1960s.[74]

We can analyze junctions between different types of semiconductors in a manner analogous to those already examined. In particular, an idealized energy-band model (Anderson model[75]) can be used that is adequate for most purposes as illustrated in Fig. 2.30 for the case of a *pn* junction between a narrow and wide band gap semiconductor, respectively. Similar comments apply to isotype heterojunctions,

[74] Shockley proposed using heterojunctions for devices in 1951 while in the same year Gubanov presented studies on their theoretical properties. In 1957, Kroemer provided important extensions to the earlier heterojunction studies and device proposals, followed by pioneering experimental studies by Anderson around 1960. Subsequently, the ability to grow different semiconductor layers with atomic precision and minimal lattice mismatch allowed high-quality heterojunction interfaces to be developed.

[75] R. L. Anderson, IBM J. Res. Dev. 4, 283 (1960). The Anderson model is similar to the idealized Schottky barrier model and can be modified to also include the effects of non-idealities such as interface states, etc.

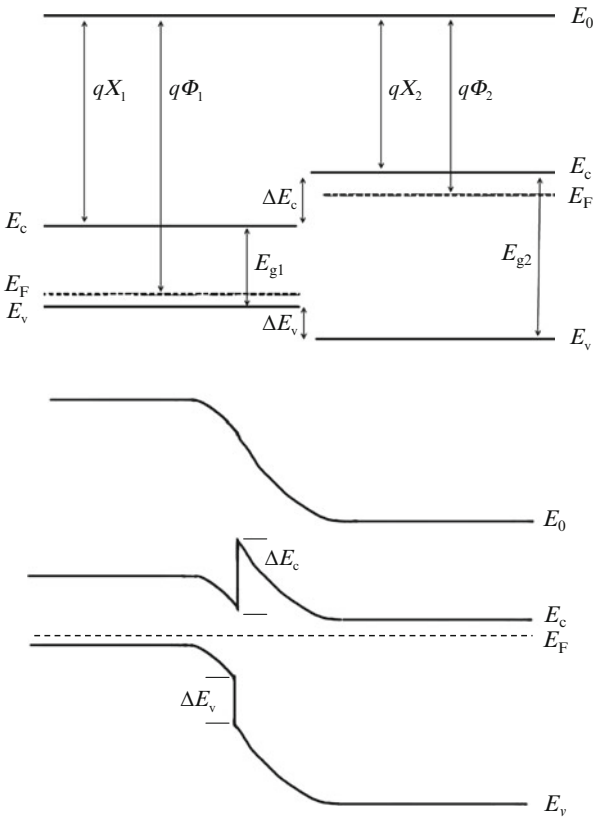


Fig. 2.30 *pn* heterojunction band edge diagram (cf. Figs. 2.2b and 2.22b)

although one must realize that the electrical properties of any type of heterojunction will depend critically on the relative magnitudes of the band gap energies, work functions, electron affinities, and applied bias. As such, heterojunctions can display *I–V* behavior that is reminiscent of diffusive transport across a *pn* junction, transport over or through a metal–semiconductor barrier, and/or anywhere in between.

2.4 Applications of Single-Junction Devices or Diodes

The most common application of a two-terminal junction device or diode is current rectification, since it has very little resistance to current flow in one direction and a very high resistance in the other. The effectiveness of a rectifier can be characterized by defining a *rectification ratio* that relates current under forward bias to that under reverse bias.

We have also seen that because of their voltage variable junction capacitance both *pn* and Schottky diodes can be employed as varactors in various applications. Varactors typically operate under reverse bias in order to maximize the voltage range that can be used to adjust the capacitance and avoid excessive currents.

Another common use of a diode is to provide *isolation*, as we have already seen for the case of reverse-biased *pn* junctions in integrated circuits. More generally, diodes are used to provide electrical isolation and protection for a wide variety of applications and systems in the form of electrostatic discharge (ESD) protection, which prevents high voltages by passing large currents through either forward-biased or reverse-biased (Zener) diodes. Other applications where diodes are used include diode logic circuits, battery/mains switching circuits and some matrix or crossbar switches.

The well-defined forward-bias characteristics of junction diodes also enable them to be used as very accurate temperature sensors. A constant current is typically passed through the diode while the forward voltage variation is measured as a function of temperature (see Problem 3). The "freezing out" of dopants at very low temperatures increases the effective resistance of the diode and causes the voltage to increase much more rapidly, which sets a lower limit for diode thermometers of about 1 or 2 K in practice. On the other hand, depending on the type of semiconductor and doping levels diode temperature sensors can operate up to 500 K or more.[76]

2.4.1 *pn* Diodes

pn junctions generally have a smaller reverse saturation current density than typical Schottky diodes. In addition, *pn* junctions are usually preferred for Zener diode applications because of the greater control that can be achieved through dopant concentrations and distributions.

A very important and distinct application of *pn* junctions is in *optoelectronics*, in particular light-emitting devices such as light-emitting diodes (LEDs) and laser diodes that rely on the emission of light via e–h pair recombination in and around the space-charge region of junctions made using direct band gap semiconductors. *pn* junctions are also commonly used for solar cells and photodiodes (the sensing elements in many digital cameras) as discussed further below. Applications in *solid-state lighting* (based on high-efficiency LED light sources) and solar energy production are some of the strongest growth areas for *pn* junction devices.

[76] As temperature increases the forward voltage drop becomes progressively smaller and this ultimately limits sensitivity at high temperatures. The increased intrinsic carrier concentration may also affect the majority carrier levels in the semiconductor at elevated temperatures (see Appendix A) and alter the expected junction behavior.

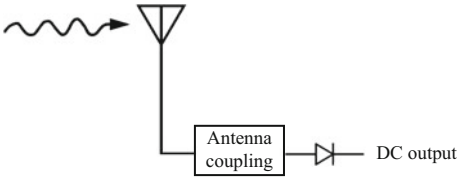


Fig. 2.31 Diode radio detection, general schematic. The incoming radiation (small signal or large signal) is rectified by the diode, which results in a nonzero dc output current component that can be used to produce an audible output or processed further (e.g., amplified, etc.), depending on the application

2.4.2 Schottky Diodes

As discussed earlier, since the Schottky diode is a majority carrier device it can be operated up to THz frequencies, which usually makes it the diode of choice for high-speed applications, such as microwave detectors, mixers, varactors, oscillators, etc. Schottky diodes are also used aboard satellites to detect changes in the Earth's atmospheric chemical species via GHz and THz gas spectroscopy.[77]

Another advantage of Schottky diodes for some applications is that they typically have a lower voltage drop (i.e., lower "turn-on" voltage) in forward bias than *pn* junctions. For example, Schottky diodes fabricated on n-type silicon will typically have a built-in potential about 200–300 mV smaller than a comparable *pn* junction. Thus a Schottky diode placed in parallel with a *pn* junction diode will prevent it from passing significant current in the forward direction or in other words the *pn* junction is "clamped" to a lower forward voltage. Such Schottky-clamped devices can be used to improve the switching speed of digital logic circuits (Chap. 3), in addition to general voltage clamping circuit and other applications where a low forward voltage drop and/or fast switching is advantageous.

Lastly, from a manufacturing point of view, Schottky diodes generally involve simpler processing steps compared to *pn* junctions. In addition, lower temperatures during fabrication may provide an economic and environmental advantage for some applications due to the smaller input energy needed and also increases process compatibility.

Panel 2.4: Microwave Detection and the Development of Semiconductors Historically, one of the most important applications of diode rectifiers has been the detection of radio signals as discussed briefly in Chap. 1 in relation to vacuum tube diodes. A basic radio detection scheme is shown in Fig. 2.31. Early telegraph and

[77] The Earth Observing System (EOS) Microwave Limb Sounder (MLS) is one prominent example, which is part of NASA's Aura satellite and uses GaAs-based Schottky diode mixers to detect radiation via heterodyning. Some images of satellite data collection results are shown at the end of this chapter.

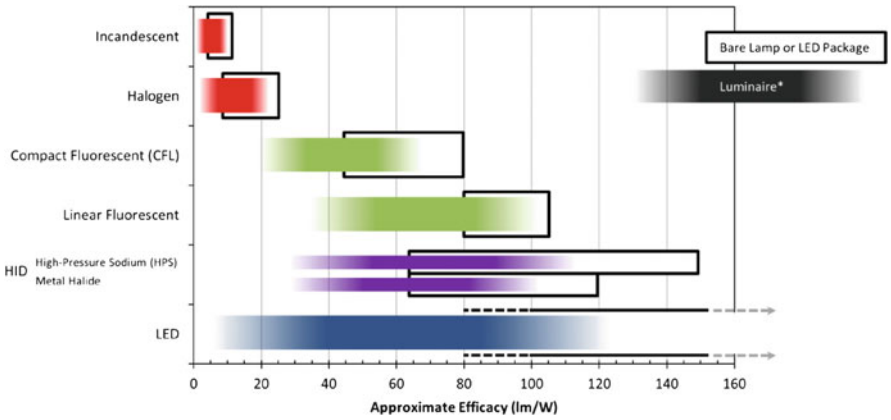


Fig. 2.32 Light output per watt for various modern electric light sources. Solid-state lighting based on LEDs continues to improve in performance and increase its market share (Luminaire refers to the entire lighting package) (Source: US Department of Energy, 2013)

radio receivers relied on point-contact diode detectors. These so-called crystal radio sets were very widespread in the early twentieth century until the development of vacuum tube technology. However, the need for microwave technology during the Second World War, in particular microwave radar, required detectors that operated at higher frequencies than earlier radio. Due to their size, typical vacuum tube devices were not able to respond fast enough to enable microwave detection. Thus, a resurgence in the use of the much faster point-contact diodes occurred. Germanium and silicon crystal rectifiers would play key roles as radar receivers in the war. Very importantly, this led to renewed interest in semiconductors following the war and paved the way for the development of modern electronics in the second half of the twentieth century.

To complete the discussion on applications we mention that isotype junctions are very useful as ohmic contacts (as alluded to above) and this is their most widespread application, particularly for planar integrated circuit technology, and heterojunctions are widely used for optoelectronics including LEDs and diode lasers because they often allow tremendous improvements in performance by enhancing the radiative recombination of carriers. For example, modern LEDs based on different types of heterojunctions have now become among the most efficient sources of light available and these set the standard for solid-state lighting applications at present (see Fig. 2.32). Heterojunctions are also often used in solar cells, which is discussed more generally in the following panel.

Panel 2.5: Photovoltaics The amount of energy being generated by the sun is immense. Nuclear fusion creates approximately 10^{20} J/s of power; most of it emitted in the form of electromagnetic radiation in the UV and IR regions of the spectrum. This power output is expected to continue for another ten billion years or so and is the source of input energy for the earth. Of the energy produced by the

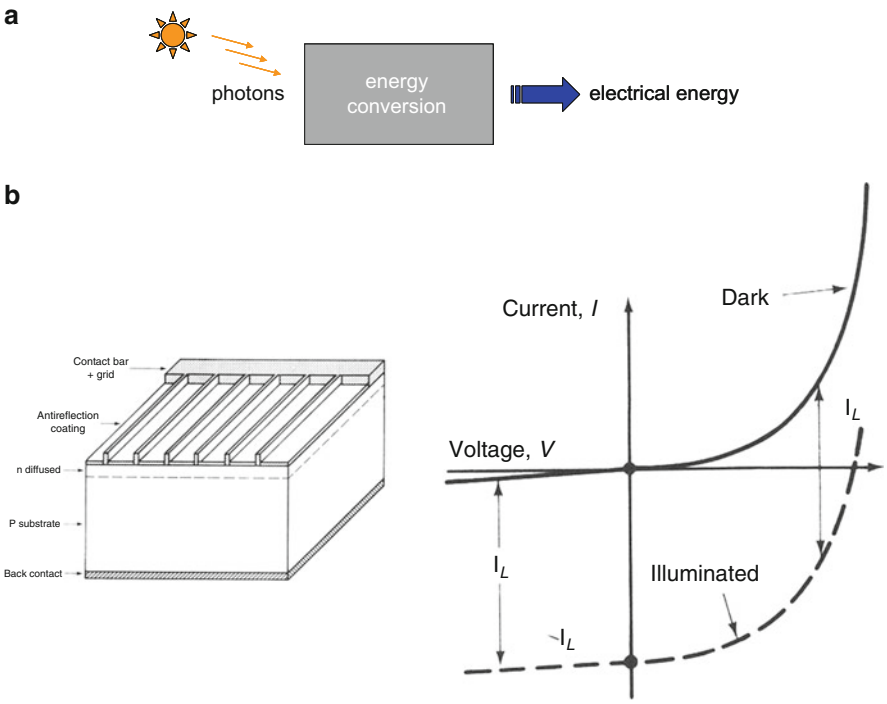


Fig. 2.33 (**a**) General concept for a solar cell that converts the energy present in sunlight into usable electrical energy. The conversion of light to electrical energy can be direct or indirect. (**b**) Illuminated diode *I–V* characteristic and simple *pn* junction solar cell device structure. The ideal diode characteristic is shifted downward by the photogenerated current I_L (After [10])

sun approximately 1,367 W/m^2 reaches the earth's atmosphere in the form of solar EM radiation (the solar constant).[78] Note that even with atmospheric losses the total amount of solar energy striking the earth's surface is approximately 10,000 times the total energy consumption of the world's entire population (~ 15 TW). Thus if this energy could be harnessed it could provide much of the world's energy needs. Over the past decade interest and growth in such solar cell technology have increased dramatically due to a combination of social, economic, and environmental factors.

The basic concept of a solar cell is shown in Fig. 2.33a. The development of practical solar cell technology has been based on finding an efficient and inexpensive means to convert the incoming solar radiation (photons) into electrical energy.

[78] This is however attenuated by the atmosphere before reaching the earth and the amount of "air mass" (AM) through which the light passes is used to describe the incident solar energy. Terrestrial solar cell performance is typically specified with respect to the AM1.5 spectrum (48° from vertical) or about 1,000 W/m^2.

In 1839 A. E. Becquerel observed that AgCl-coated Pt electrodes in an electrochemical cell (a semiconductor–liquid junction) produced a voltage when exposed to light. Subsequently, similar behavior was observed in many other materials, particularly solid-state junctions, and this phenomenon is now generally referred to as the *photovoltaic effect*. Virtually all solar cell technology currently in use is based on the photovoltaic effect, and thus these devices are also known as solar photovoltaic cells. Upon illumination solar cells can deliver power to an external circuit in a direct and clean manner (i.e., essentially nonpolluting).

The most common solar cells and those developed first commercially are based on *pn* junctions (see Fig. 2.33b for a basic solar cell structure). When the junction is illuminated with photons near or above the band gap energy the excess electron–hole pairs that are created will be swept out of the depletion region by the built-in electric field and result in current flow, i.e., *photovoltaic energy conversion* will occur. By extending our previous "dark" treatment of the *pn* junction we can show that the illuminated junction *I–V* characteristic is essentially a modification of the ideal diode equation as shown in Fig. 2.33b.

We can proceed in an identical manner to the previous ideal diode analysis; however, an extra term is now added to the diffusion equations that represents the generation of electron–hole pairs due to the incident light:

$$\begin{aligned}\frac{\partial n'}{\partial t} &= D_n \frac{\partial^2 n'}{\partial x^2} - \frac{n'}{\tau_n} + G \\ \frac{\partial p'}{\partial t} &= D_p \frac{\partial^2 p'}{\partial x^2} - \frac{p'}{\tau_p} + G\end{aligned} \tag{2.80}$$

As before, we can now look for the steady-state solutions to these equations. For simplicity, we assume that the junction is uniformly illuminated with photons, i.e., G is a constant. For steady state in the n-side of a *pn* junction we then have

$$0 = D_p \frac{d^2 p'_{\mathrm{n}}}{dx^2} - \frac{p'_{\mathrm{n}}}{\tau_p} + G \tag{2.81}$$

This is similar to the dark junction case, but the addition of the generation term has now resulted in a *nonhomogeneous* (linear) differential equation. Recall that a general solution to such an equation consists of the solution to the homogeneous equation (that we obtained earlier) plus any particular solution. We therefore get the following solution:

$$p'_{\mathrm{n}}(x) = A \exp\left(-\frac{x - x_{\mathrm{n}}}{L_p}\right) + B \exp\left(\frac{x - x_{\mathrm{n}}}{L_p}\right) + G\tau_p \tag{2.82}$$

where once again A and B are constants and L_p is the hole diffusion length (and L_n is the analogous electron diffusion length). To find the constants we

can again consider two limiting cases based on the length W_B of the n-region from the junction to the ohmic contact (see Fig. 2.4). For the long-base diode the solution turns out to be[79]

$$p'_n(x) = \left[p_{n0}\left(e^{qV_a/k_BT} - 1\right) - G\tau_p\right]\exp\left(-\frac{x - x_n}{L_p}\right) + G\tau_p \tag{2.83}$$

Note that, as expected, far away from the junction the excess carrier concentration approaches a constant value determined by the generation rate and the carrier lifetime. Using the above expression we can now calculate the diffusion current density due to holes at the edge of the space-charge region:

$$\begin{aligned} J_p(x) &= -qD_p\frac{dp_n}{dx} \\ &= qD_p\frac{p_{n0}}{L_p}\left(e^{qV_a/k_BT} - 1\right)\exp\left(-\frac{x - x_n}{L_p}\right) - qGL_p\exp\left(-\frac{x - x_n}{L_p}\right) \end{aligned} \tag{2.83}$$

The total current flowing through the *pn* junction is obtained as before by summing the two minority carrier currents flowing across the junction (electrons and holes) giving

$$\begin{aligned} J &= J_p(x_n) + J_n(-x_p) \\ &= qn_i^2\left(\frac{D_p}{N_dL_p} + \frac{D_n}{N_aL_n}\right)\left(e^{qV_a/k_BT} - 1\right) - qG(L_p + L_n) \\ &= J_0\left(e^{qV_a/k_BT} - 1\right) - qG(L_p + L_n) \end{aligned} \tag{2.84}$$

That is, the ideal diode equation (dark) current is simply reduced or shifted by the photogenerated current. Thus, the photovoltaic effect creates a current that flows in the reverse direction through the junction.

A contribution that is not included in the above derivation is photo generation of carriers within the depletion region. We can easily include this additional current by adding another term to the photogenerated current proportional to the depletion width, x_d:

$$J = J_0\left(e^{qV_a/k_BT} - 1\right) - qG(L_p + L_n + x_d) \tag{2.85}$$

[79] Note the excess carrier concentration at the edge of the depletion region will be determined (for the case of low-level injection, as before) by the voltage appearing across the junction, V_a (cf., Eq. (2.23)). (In the case of an illuminated junction V_a is not strictly an applied bias, but we keep the same notation for simplicity.)

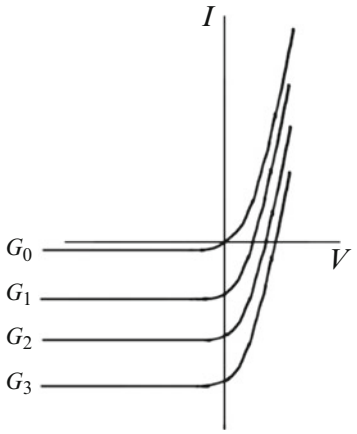


Fig. 2.34 *pn* junction *I–V* characteristics for increasing levels of illumination denoted by optical electron–hole pair generation rate *G*

Note that this extra term can be ignored if it is small compared to the minority carrier diffusion lengths. The above equation also indicates that the photogenerated current is caused by carriers generated within a diffusion length of the depletion region. The diffusion lengths, along with the depletion region itself, define the active "collection regions" of a *pn* junction solar cell.

For the short-base diode, we can repeat the above analysis; however, recall that the net result was simply a modification of the reverse saturation current density term appearing in the ideal diode equation, and other than replacing the diffusion lengths with the physical dimensions of the neutral regions, the functional form of the solution will be identical to that above.

Our (idealized) analysis of the illuminated *pn* junction has essentially shown that the current is given by

$$I = I_{\text{dark}} - I_L \tag{2.86}$$

where I_{dark} is the usual ideal diode equation current and I_L is the current due to optical generation:

$$I = I_0\left(e^{qV_a/k_BT} - 1\right) - qAG\left(L_p + L_n + x_d\right) \tag{2.87}$$

To first order this means that the illuminated *I–V* characteristics are identical to the dark characteristics except that they are translated downward by I_L as illustrated in Fig. 2.34. The point at which the curves cross the current axis is known as the *short-circuit current*, I_{sc} (i.e., when $V_a = 0$), which has a magnitude equal to I_L. On the other hand, when there is an open circuit across the junction device, $I = 0$ and the open-circuit voltage is

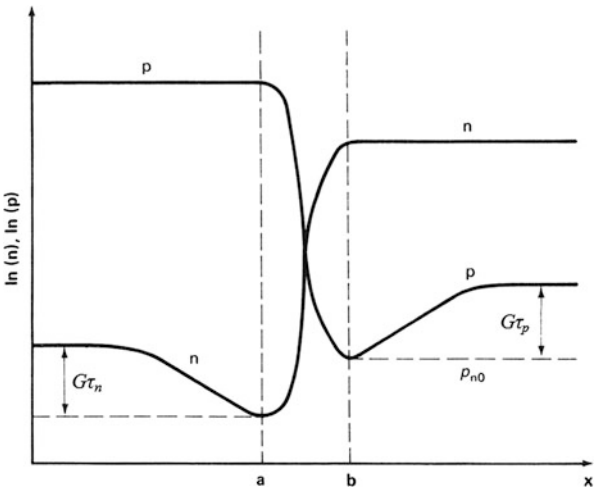


Fig. 2.35 *pn* junction short-circuit current caused by minority carriers diffusing towards the space-charge region during steady-state illumination (Adapted from [10])

$$V_{\text{oc}} = \frac{k_{\text{B}}T}{q} \ln\left(\frac{I_L}{I_0} + 1\right) \tag{2.88}$$

Thus, V_{oc} depends on the optical generation rate through I_L and the properties of the semiconductor through I_0. Recall that V_{oc} will never be larger than the built-in potential of the junction. Depending on the intended application, an illuminated *pn* junction can be operated in different quadrants of its *I–V* characteristic. For example, the third quadrant is typically used for photodetection and the device is then a *photodiode*. For the purpose of power/energy generation we are most interested in the *fourth quadrant* since in this case power can be extracted from the device (similar to a battery).

It is useful to briefly discuss the physical mechanisms of the voltage appearing across an illuminated junction: When light shines on the *pn* junction, one without an external bias voltage, each absorbed photon creates an e–h pair. When these carriers diffuse to the junction (or are created within the depletion layer) the built-in electric field separates them. This separation of charge produces a forward voltage across the barrier since the electric field of the photoexcited carriers is opposite to the built-in field (cf. Fig. 2.2b). As stated previously this is the origin of the photovoltaic effect and causes the appearance of the open-circuit voltage defined above for an illuminated junction.

We can also obtain some insight by examining the distribution of carriers near the illuminated junction when it is short-circuited as shown in Fig. 2.35. The origin of the short-circuit current can now be seen as due to the diffusion of carriers from high concentration (away from the junction) to low concentration (towards the junction) where these are swept away by the built-in electric field. The concentration gradient is maintained by the photogeneration of carriers. In steady state,

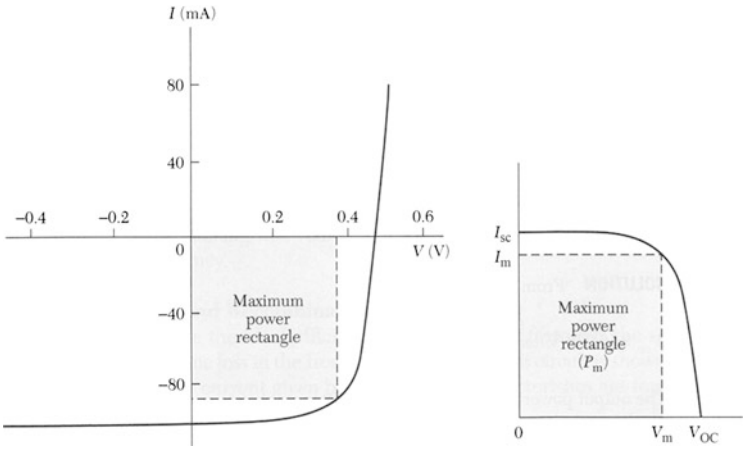


Fig. 2.36 Solar cell maximum power rectangle superimposed on diode *I–V* characteristic. By convention the data is often flipped vertically in order to give positive values as shown by the diagram on the right-hand side (After [2])

the illuminated junction will in general be somewhere in between the two extremes of short- and open-circuit behavior as the system responds to the nonequilibrium condition created by the exposure to light. In a sense, the response of the illuminated junction to a bias is opposite to that of the "dark" junction (first quadrant) since the output current magnitude now decreases with bias instead of increasing.

The output power for any operating point in the fourth quadrant can be found from the usual expression

$$P = IV = I_0 V\left(e^{qV_a/k_BT} - 1\right) - I_L V \tag{2.89}$$

and the condition for maximum power output is obtained when $dP/dV = 0$ or

$$V_m = \frac{k_B T}{q}\ln\left[\frac{1 + (I_L/I_0)}{1 + (qV_m/k_B T)}\right] = V_{oc} - \frac{k_B T}{q}\ln\left(1 + \frac{qV_m}{k_B T}\right) \tag{2.90}$$

which along with the corresponding current I_m determines the maximum power output P_m. These parameters define the *maximum power rectangle* shown in Fig. 2.36. We can see that the area of the rectangle will always be less than the product of the short-circuit current and open-circuit voltage. The ratio

$$\mathrm{FF} = \frac{I_m V_m}{I_{sc} V_{oc}} \tag{2.91}$$

is known as the *fill factor* and is an important figure of merit for solar cell design. Lastly, the *power conversion efficiency* provides an overall measure of solar cell performance:

$$\eta \equiv \frac{P_{\mathrm{m}}}{P_{\mathrm{in}}} = \frac{I_{\mathrm{m}} V_{\mathrm{m}}}{P_{\mathrm{in}}} = \frac{\mathrm{FF} I_{\mathrm{sc}} V_{\mathrm{oc}}}{P_{\mathrm{in}}} \tag{2.92}$$[80]

where P_{in} is the incident optical energy per unit time or the input power. Thus, to maximize efficiency all three solar cell parameters—fill factor, short-circuit current, and open-circuit voltage—should be maximized.[81]

Photovoltaics currently only provide a very small portion of the world's energy: Approximately 100 GW of solar power is installed worldwide. At present most of these installations are based on silicon (~80 %), although other materials are being actively developed as well. The photovoltaic market has been growing annually at a rate of approximately 50 %. New installations currently exceed 20 GW annually. If these types of growth rates continue terawatt levels of power could be reached by 2030. Many alternative materials and technologies are currently being developed to provide efficient solar energy conversion with lower production costs. In addition, these emerging technologies may allow implementations not readily achievable with existing devices (e.g., flexible panels, access to different regions of the electromagnetic spectrum, enhanced stability, integration with other devices and materials, etc.).

Although the material on junctions in this chapter has been somewhat extensive, it is worth reemphasizing the point made at the beginning of the chapter that this is because such junctions are the basis of virtually all modern electronics: Being able to control the distribution of charges in a semiconductor to create junctions has allowed the development of solid-state devices and circuits that can perform complex tasks,[82] which are the foundation of today's information-based society, as will be discussed further in Chaps. 3 and 4.

References

1. Shockley, W.: The Theory of $p-n$ Junctions in Semiconductors and $p-n$ Junction Transistors. Bell Syst. Tech. J. **28**, 435 (1949)
2. Sze, S.M., Ng, K.K.: Physics of Semiconductor Devices, 3rd edn. Wiley Interscience, Hoboken (2007)
3. Kingston, R.H.: Switching Time in Juction Diodes and Junction Transistors. Proc. IRE **42**, 829 (1954)

[80] Not to be confused with the diode ideality factor.

[81] Sources of solar cell losses include the inability to absorb photons with energy less than the band gap, heat generated by large energy photon absorption, reflection losses, carrier recombination, and parasitic resistances.

[82] Previously this was only possible in other areas of technology, for example, using the potential energy of a spring or the pressure/temperature differential of gases or of a chemical reaction to perform a useful task. Solid-state electronics has far surpassed the complexity of any other type of man-made "machine" in terms of the number of working parts operating together, as exemplified by the integrated circuit.

4. Muller, R.S., Kamins, T.I.: Device Electronics for Integrated Circuits, 3rd edn. Wiley, New York (2003)
5. Grove, A.S.: Physics and Technology of Semiconductor Devices. Wiley, New York (1967)
6. Shur, M.S.: Introduction to Electronic Devices. Wiley, New York (1995)
7. Neudeck, G.W.: The PN Junction Diode, 2nd edn. Prentice-Hall, Boston (1989)
8. Smith, R.A.: Semiconductors, 2nd edn. Cambridge University Press, Cambridge (1978)
9. Neamen, D.A.: Semiconductor Physics and Devices, 3rd edn. McGraw-Hill, New York (2003)
10. Green, M.A.: Solar Cells. Prentice-Hall, Upper Saddle River (1982)

Problems

1. *Long- vs. short-base diodes*. An ideal silicon *pn* diode is formed by diffusing a high concentration of phosphorus into a 75-μm-thick boron-doped wafer having a resistivity of 5 Ω-cm and minority carrier lifetime of 5 μs. The junction is formed 10 μm below the surface with area 10^{-4} cm^2. (1) Find the built-in potential. (2) Calculate the current flowing through the diode under an applied forward bias of 0.5 V. (3) Is I_0 for an ideal diode always constant?
2. *Ohmic voltage drops*. Consider a silicon short-base diode with the following parameters:

$$N_d = 2 \times 10^{17}\text{cm}^{-3} \text{ and } N_a = 5 \times 10^{18}\text{cm}^{-3}$$
$$x_B = x_E = 5\ \mu m \text{ and } A = 10^{-4}\text{cm}^2$$

   Find the voltage dropped in the neutral regions of the diode for an applied forward bias of 0.7 V.
3. [83] *Diode temperature dependence*. An ideal long-base Si *pn* diode has $N_a = 10^{17}$ cm^{-3}, $N_d = 7 \times 10^{16}$cm^{-3}, and a cross-sectional area of 10^{-3}cm^2. (1) If $\tau_n = \tau_p = 1$ μs, calculate the current flowing through the junction under an applied bias of 0.5 V. Repeat your calculation for a temperature of 500 K. (2) Sketch the thermal equilibrium band edge diagram of the *pn* junction for both temperatures. (3) If the junction in this question were required to absorb light, at what wavelength would it begin to absorb strongly?
4. *pn diode storage time*. Compare the accuracy of Eqs. (2.61) and (2.62b) to the full solution of the continuity equation given by Eq. (2.62a), for I_R/I_F ranging from 0.01 to 100.

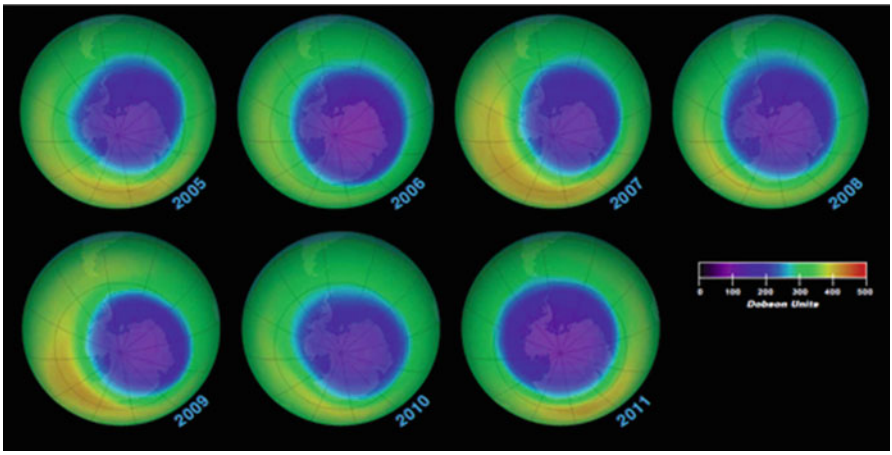
[83] This problem may be somewhat difficult and/or lengthy.

5. *Metal–semiconductor band edge diagrams*. An ideal metal–semiconductor junction is formed between platinum (work function 5.3 eV) and p-type silicon. What type of contact results?

Microwave Limb Sounder onboard the Aura Satellite

The MLS (located near the foreground in the image below) observes microwave emission from gas molecules (e.g., O_3, H_2O, CO, SO_2) from 118 GHz to 2.5 THz using GaAs-based Schottky diodes (solar panels based on silicon *pn* junctions power the satellite as well). The data shown is based on annual MLS Earth ozone concentration measurements taken over the south pole.

Source: JPL/NASA


2005
2006
2007
2008
2009
2010
2011
0
100
200
300
400
500
Dobson Units

Chapter 3
Bipolar Transistors

"This thing's got gain!"

W. H. Brattain, *Bell Laboratories 1947*

Amplifying a signal typically requires an intervening effect or party as illustrated by the analogy in Fig. 3.1a. Similar comments apply to switching (Fig. 3.1b). In the early twentieth century vacuum tube triodes[1] (Fig. 3.1c) began replacing the earlier mechanical and electromechanical amplifying/switching devices, as discussed in Chap. 1. Such triodes were the first all-electronic amplifiers and switches.

During the first half of the twentieth century various workers began a quest to demonstrate a solid-state version of the triode, i.e., a solid-state electronic amplifier[2] or switch that could be much smaller and more reliable than vacuum tubes. This proved to be a very difficult task in practice; however by the late 1940s the semiconductor research group at Bell Labs led by Shockley[3] was able to demonstrate the first semiconductor triode, dubbed the *transistor*, in the form of a point-contact structure that was created by Brattain and Bardeen[4] (Fig. 3.2). This initial invention was followed shortly thereafter by the more practical bipolar transistor,[5] which we discuss in detail below.

[1] The term triode follows the naming convention of diodes and here refers to the three different paths in such devices. Tetrode and pentode, etc., vacuum tubes were also developed, containing 4 and 5 terminals, etc., respectively.

[2] Important early work using point-contact devices and circuits for radio signal detection and amplification was due to Eccles and Losev (often spelt Lossev).

[3] A large part of this early effort was devoted to developing field effect devices, which, although unsuccessful at the time, allowed important insights that led to the first transistors. Field effect transistors are covered in Chap. 4.

[4] J. Bardeen and W. H. Brattain, Phys. Rev. **74**, 230 (1948).

[5] W. Shockley, Bell Syst. Tech. J. **28**, 435 (1949).

C. Papadopoulos, *Solid-State Electronic Devices: An Introduction*,
Undergraduate Lecture Notes in Physics, DOI 10.1007/978-1-4614-8836-1_3,

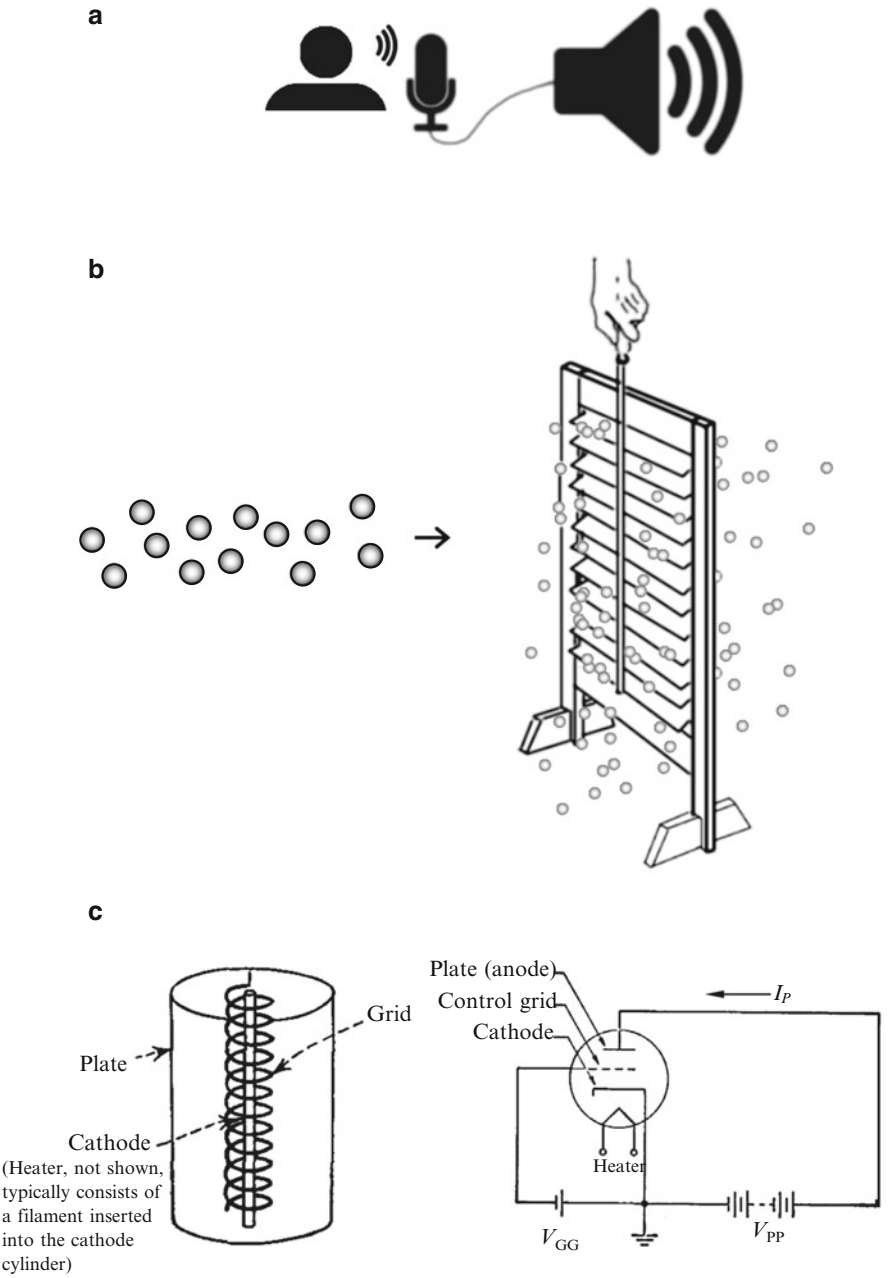


Fig. 3.1 Amplification and switching analogies. (**a**) Mechanical vibration (sound) impinging on a microphone is used to amplify the signal via a speaker. (**b**) A stream of particles is switched on or off by the opening or closing of the slats in a shutter (Adapted from "How the Repeater Repeats", Western Electric News, March, 1917.) (**c**) The vacuum tube triode was the first solely electronic implementation of an amplifier or switch. Electrons emitted from the heated cathode will reach the

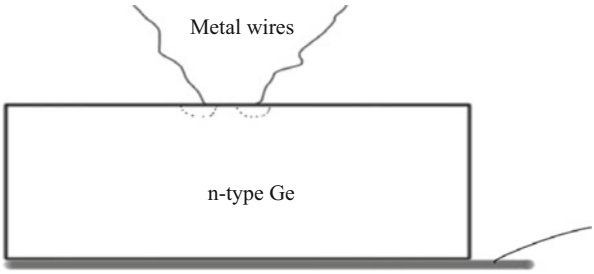


Fig. 3.2 Original point-contact transistor device structure schematic. Two contacts are made in close proximity to each other by pressing metallic wires onto the semiconductor surface. The current through one of the contacts could be modulated by the other (Contact to the semiconductor was made using a broad area electrode on the bottom of the crystal as shown)

The first transistor and subsequent developments are considered a defining point in the history and development of electronics. The dominance of solid-state electronic devices and integrated circuits in modern technology is directly linked to the initial semiconductor triode research.

3.1 Transistor Effect

In Chap. 2, we saw that a *pn* junction biased so that the p-region is positive with respect to the n-region conducts current because holes are injected across the junction from the p-region and electrons are injected from the n-region, both of which have plentiful supplies of majority carriers. Consequently, current increases rapidly as voltage increases and the barrier at the junction decreases.

Currents are much smaller under reverse bias because they are carried only by minority carriers generated either in the space-charge region or nearby. However, if the supply of minority carriers in the vicinity of the junction could be somehow enhanced the current passed by a reverse-biased junction would increase. For example, this can result from radiation incident on the diode with energy greater than the band gap (as in a photodiode), which can be considered an optical analogue of the microphone.

Another means of enhancing the minority carrier population in the vicinity of a reverse-biased *pn* junction is to locate a forward-biased *pn* junction very close to it:

Fig. 3.1 (continued) anode plate depending on the bias of the intervening control grid. For negative biases on the grid with respect to the cathode electrons will be repelled and increasingly prevented from reaching the anode. The plate current is therefore modulated by the control grid. (Triode sketch adapted from D. G. Fink, *Engineering Electronics*, McGraw-Hill, New York, 1938) (Circuit schematic adapted from *Principles of Electricity applied to Telephone and Telegraph Work*, American Telegraph and Telephone Company, 1953)

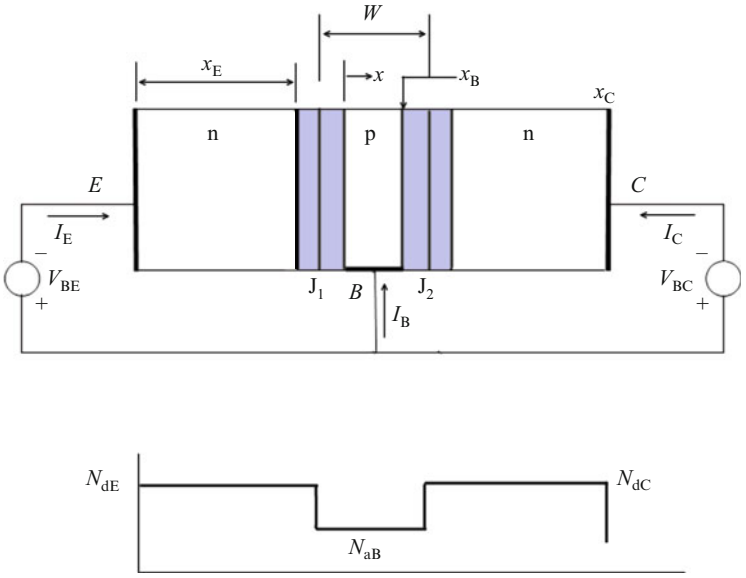


Fig. 3.3 Idealized or prototype *npn* bipolar transistor structure. The idealized structure has constant doping levels and is symmetric from emitter to collector as illustrated. Similar comments apply for the analogous *pnp* structure

The essential principle of the bipolar[6] junction transistor (BJT) is the controlled injection of minority carriers into a reverse-biased *pn* junction from a second nearby forward-biased *pn* junction. We can therefore modulate the current flow in one junction by changing the bias on another.

We consider a simple idealized, or prototype, *npn* transistor[7] structure to illustrate the basic principles of a BJT as depicted in Fig. 3.3. The two *pn* junctions are spaced a distance W apart in a single bar of semiconductor material. The bar has a uniform cross-sectional area A and the junctions are located close enough to one another so that electrons injected across junction J_1 when V_{BE} is positive reach the vicinity of junction J_2, i.e., the junction spacing is small enough that few electrons are lost by recombination in the middle p-region.

The middle region is called the *base* of the transistor, the n-type region adjacent to the injecting (or emitting) junction is called the *emitter*, and the n-type region adjacent to the collecting junction is called the *collector*.

[6] The term "bipolar" is used to describe devices that depend on both types of carriers (e.g., the *pn* diode or BJT). "Unipolar" devices on the other hand usually refer to devices that mainly rely on majority carriers for their operation (e.g., Schottky diodes and MOSFETs).

[7] Although most of the discussion in this chapter focuses on *npn* transistors, the corresponding equations for *pnp* transistors will have the same form but with the current directions and voltage polarities reversed.

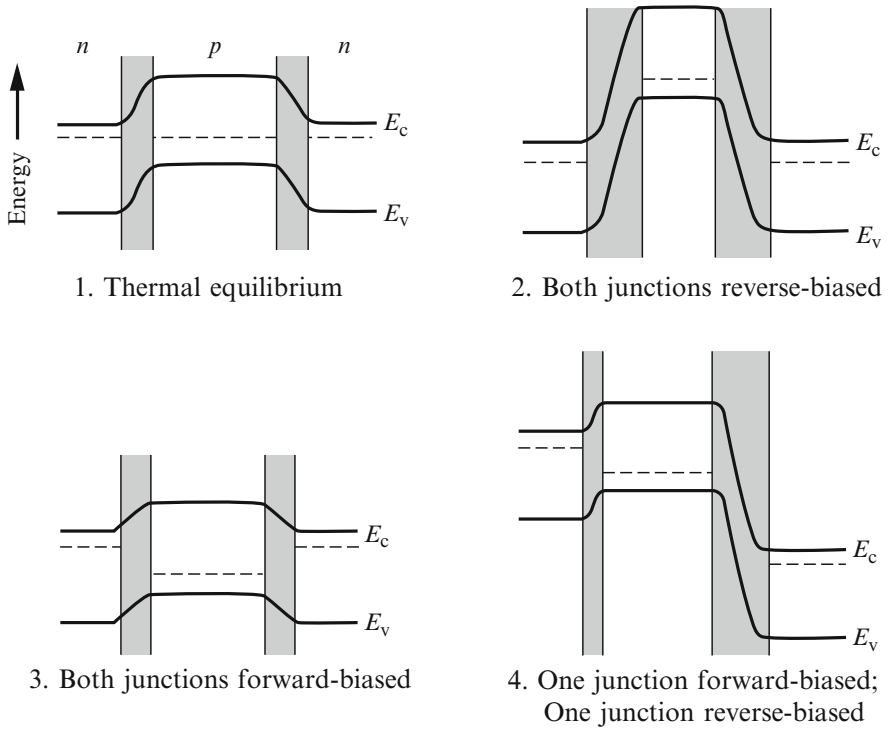


Fig. 3.4 *npn* transistor band edge diagrams for different biasing configurations indicated

Before proceeding with further analysis, we can use the prototype structure to illustrate the basic bipolar transistor properties by examining the *npn* band edge diagrams for various combinations of junction biases (Fig. 3.4):

1. Thermal equilibrium
   With zero applied bias there are few electrons in the base. The energy barriers prevent electrons in the n-type regions from traveling across the base.
2. Both junctions reverse biased
   Negative bias applied to both junctions increases the barriers and depletes the base region of the few electrons that were present at thermal equilibrium.
3. Both junctions forward biased
   Positive bias applied to both junctions injects electrons into the base region by lowering the built-in barriers. The resulting large increase in the base electron density leads to electrons flowing readily between the two junctions.
4. One junction forward biased and one junction reverse biased
   This configuration is particularly important for device applications in signal amplification. The forward-biased emitter junction injects electrons into the base, which the reverse-biased collector junction sweeps out of the base and into the *n*-type collector region. This is known as the (forward) *active bias* condition.

The resulting excess electron distribution in the base under active bias follows from our analysis of the short-base *pn* diode (recombination is negligible; constant doping density). The electron density at the emitter edge of the base region depends exponentially on the voltage V_{BE}, while at the collector edge of the base the electron density is negligible. We can therefore write the excess electron concentration for active bias in the base of the prototype BJT geometry as

$$n'_p = n_{p0}\left[e^{qV_{BE}/k_BT}\left(1-\frac{x}{x_B}\right)-1\right], 0 \le x \le x_B \tag{3.1}$$

and the electron current density for V_{BE} much greater than k_BT/q (usually the case for active bias) is therefore

$$J_n = qD_n\frac{dn_p}{dx} = \frac{-qD_n n_i^2 \exp(qV_{BE}/k_BT)}{N_{aB}x_B} \tag{3.2}$$

Thus in the active bias configuration when V_{BC} is zero or negative and V_{BE} is substantially larger than k_BT/q an electron current,

$$J_n \approx -J_0 \exp\left(\frac{qV_{BE}}{k_BT}\right) \tag{3.3a}$$

flows from left to right across the collecting junction (junction 2 in Fig. 3.3). If we use the standard convention that *currents flowing into a transistor are positive*, J_n is equal to $+J_C$, a positive collector current density, i.e.,

$$J_C \approx J_0 \ \exp\left(\frac{qV_{BE}}{k_BT}\right) \tag{3.3b}$$

This equation shows explicitly that under active bias the collector current depends exponentially on emitter–base voltage.

We are interested in learning how electrons are injected from the emitter into the base, flow across the base, and reach the collector: First note that in a bipolar transistor structure there is negligible flow of holes (base majority carriers) from junction 1 to junction 2 or vice versa. This is true under all bias conditions because the flow of holes from either n-region into the p-type base must be very small. If we further assume that electron recombination in the base is small, the hole current density in the longitudinal or x direction can be written:

$$J_p = 0 = q\mu_p pE_x - qD_p\frac{dp}{dx} \tag{3.4}$$

and

$$\begin{aligned} E_x &= \frac{D_p}{\mu_p}\frac{1}{p}\frac{dp}{dx} \\ &= \frac{k_BT}{q}\frac{1}{p}\frac{dp}{dx} \end{aligned} \tag{3.5}$$

Thus, the condition of zero hole current in the base leads to an equation describing the longitudinal electric field. This (built-in) field[8] depends on the magnitude of the hole density in the base and on its gradient.

On the other hand, the *electron* current density flowing between the junctions can be quite significant if one of the junctions is forward biased:

$$J_n = q\mu_n n E_x + qD_n \frac{dn}{dx} \tag{3.6}$$

Substituting the expression for the longitudinal electric field found above gives

$$J_n = k_B T \mu_n \frac{n}{p}\frac{dp}{dx} + qD_n \frac{dn}{dx} \tag{3.7}$$

For our basic prototype transistor structure with abrupt junctions and constant base doping the electron current density reduces to

$$J_n = qD_n \frac{dn}{dx} \tag{3.8}$$

and since we are assuming the recombination of electrons in the base is small, n will vary linearly across the base (just as it did in the short-base pn diode from Chap. 2):

$$\frac{dn}{dx} = \frac{n_p(x_B) - n_p(0)}{x_B} = \frac{n'_p(x_B) - n'_p(0)}{x_B} \tag{3.9}$$

Using our previous results from the pn diode analysis we can substitute expressions for the minority carrier concentrations at the edges of the base as a function of applied bias to obtain the electron current density:

$$J_n = \frac{qD_n n_i^2}{x_B N_{aB}}\left[\exp\left(\frac{qV_{BC}}{k_B T}\right) - \exp\left(\frac{qV_{BE}}{k_B T}\right)\right] \tag{3.10}$$

This equation expresses mathematically how the current is controlled by the junction voltages. If the doping in the base is not constant (which is typically the case in integrated circuits) the expression above can be shown to take the more general form[9]:

[8] The doping gradient means that the band edges must be sloped in thermal equilibrium, which leads to the built-in electric field described by Eq. (3.5).

[9] This result arises by integrating both sides of Eq. (3.7) and using the junction law, Eq. (2.44).

$$J_n = \frac{qn_i^2\left[\exp\left(\frac{qV_{BC}}{k_BT}\right) - \exp\left(\frac{qV_{BE}}{k_BT}\right)\right]}{\int_0^{x_B}\frac{pdx}{D_n}} \tag{3.11}$$

Since the diffusion constant D_n is usually a fairly weak function of position in the base it can be replaced by an average value ($\widetilde{D}_n$) and thus removed from the integral in the denominator. With this modification the integral is just the total majority carrier density per unit area in the base, whose charge is

$$Q_B = q\int_0^{x_B} p dx \tag{3.12}$$

We can now rewrite the electron current density flowing from the first junction to the second as

$$J_n = J_0\left[\exp\left(\frac{qV_{BC}}{k_BT}\right) - \exp\left(\frac{qV_{BE}}{k_BT}\right)\right] \tag{3.13a}$$

where

$$J_0 = \frac{q^2n_i^2\widetilde{D}_n}{Q_B} \tag{3.13b}$$

Thus the equation representing the physics of transistor action in a bipolar transistor is not an explicit function of the base doping profile but depends only on the integrated base majority charge.

The result of Eq. (3.2) for forward-active bias found earlier can also be derived from the more general Eqs. (3.13a) and (3.13b) by noting that

$$Q_B = qN_{aB}x_B$$

and dropping the negligible term exp (qV_{BC}/k_BT).

Panel 3.1: Base Charges and Gummel Number The value of J_0 in Eq. (3.13) can be found experimentally from the intercept of a collector current versus emitter–base voltage plot (Fig. 3.5) and this value can be used to determine the built-in hole base charge since, from Eq. (3.13b),

$$Q_B = \frac{q^2n_i^2\widetilde{D}_n}{J_0}$$

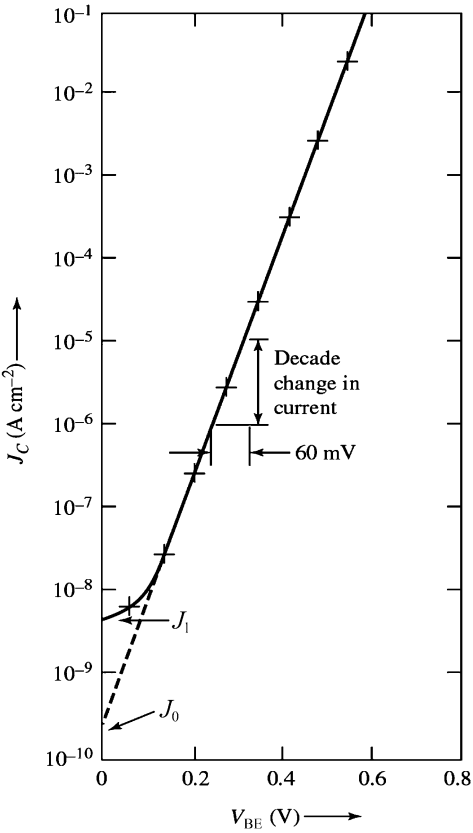


Fig. 3.5 Experimental *npn* transistor collector current density versus base–emitter voltage. J_1 represents the collector "leakage" current under reverse bias (After [1])

and all other parameters are known. This charge is determined during processing of the transistor, and the number of base-dopant atoms (per unit area) is sometimes referred to as the *Gummel number (GN)*:

$$GN = \int_0^{x_B} N_{aB}(x)dx = \frac{Q_B}{q} = \frac{qn_i^2\widetilde{D}_n}{J_0} \tag{3.14}$$

Thus the lower the total base doping level or Gummel number is, the higher the transistor current for a given bias. For high-gain transistors the Gummel number must be kept low and carefully controlled during production in order to ensure adequate performance.

3.2 Gain and Switching

3.2.1 Current Gain

Thus far our discussion of the BJT has only considered electrons flowing between the emitter and the collector—this represents the output current in an active-biased transistor. We saw that collector current is an exponential function of base–emitter voltage because forward bias on the base–emitter junction causes electron injection into the base to vary exponentially. Under active bias, these electrons are collected efficiently by the field in the base–collector space-charge region. The base–emitter terminals are thus the control electrodes for the collector current under active bias: The smaller the current that flows through the base terminal for a given positive V_{BE}, the more effective is the transistor as an amplifier because the input power (the product of V_{BE} and the base emitter current) is lower.

Several mechanisms can contribute to base–emitter current in an active-biased transistor, the main ones being:

1. Recombination of injected electrons with majority carrier holes in the neutral base region.
2. Hole injection into the emitter across the forward-biased base–emitter junction.
3. Recombination in the base–emitter space-charge region.

For the transistor to amplify effectively all of these current components should be much smaller than the collector current as illustrated schematically in Fig. 3.6. Under typical operating conditions for silicon bipolar transistors the recombination of injected electrons in the space–charge region is usually smaller than the other components shown.

The recombination of excess minority carriers in the base is directly proportional to their density and so the total base-region recombination current can be found from

$$I_{\mathrm{rB}} = qA_{\mathrm{E}} \int_0^{x_{\mathrm{B}}} \frac{\left[n - \left(n_{\mathrm{i}}^2 / N_{\mathrm{aB}}\right)\right] dx}{\tau_n} \tag{3.15}$$[10]

where A_{E} is the area of the emitter. To make I_{rB} as small as possible at a given bias (and thus increase the gain), the lifetime of electrons in the base τ_n should be maximized and the base width x_{B} should be minimized. Under active bias the injected electron density is much larger than the equilibrium density over most of the base. Also, the electron lifetime does not depend strongly on x. We can therefore write

[10] Cf. footnote of Eq. (2.37), where U is given by the integrand in Eq. (3.15).

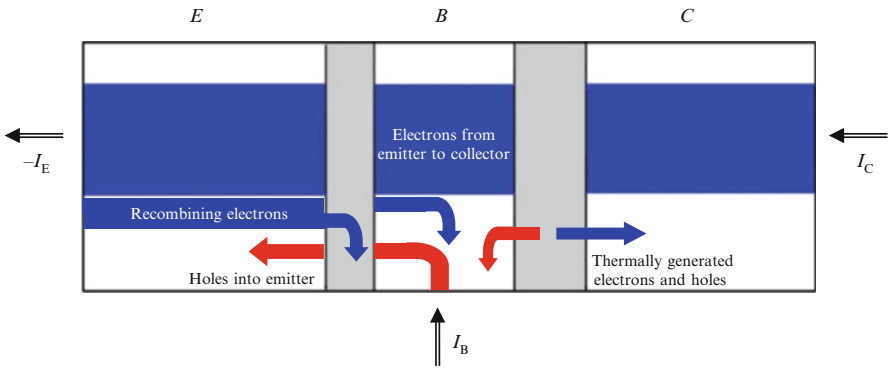


Fig. 3.6 Current components in an *npn* bipolar transistor under forward-active bias (electrons in *blue*, holes in *red*)

$$I_{rB} = \frac{qA_E}{\tau_n}\int_0^{x_B} n' dx \tag{3.16}$$

For the case of a transistor with a uniformly doped base we saw earlier that n' depends linearly on x and so the integration is easily carried through to obtain

$$I_{rB} = \frac{qA_E n_i^2 x_B}{2N_{aB}\tau_n}\left[\exp\left(\frac{qV_{BE}}{k_B T}\right) - 1\right] \tag{3.17}$$

The loss of carriers to recombination in the base is measured by the *base transport factor*:

$$\alpha_T = \frac{|I_{nC}|}{|I_{nE}|} = \frac{|I_{nE}| - |I_{rB}|}{|I_{nE}|} = 1 - \left|\frac{I_{rB}}{I_{nE}}\right| \tag{3.18}$$

where I_{nE} is the electron current injected from the emitter. Now, using the explicit expression for recombination current together with the result for injected electron emitter current (Eqs. (3.17) and (3.10), respectively), we obtain

$$\alpha_T = 1 - \frac{x_B^2}{2D_n\tau_n} = 1 - \frac{x_B^2}{2L_n^2} \tag{3.19}$$

This equation is not directly applicable to planar IC transistors because the base doping in ICs is usually graded due to dopant diffusion technology, which improves the base transport factor [see Eq. (3.13b)]. However, α_T calculated from the above equation can still be used to determine the "worst-case" transport factor in planar transistors (and is usually not the main limiting factor for transistor performance).

The injection of base majority carriers (holes) into the emitter is the main contribution to base current in modern transistors. Expressions for this current have already been derived (see Chap. 2) since the base–emitter junction under active bias

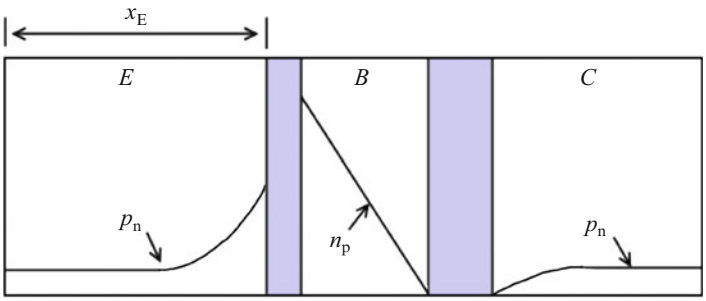


Fig. 3.7 *npn* transistor minority carrier distributions under forward-active bias for the case of a long emitter

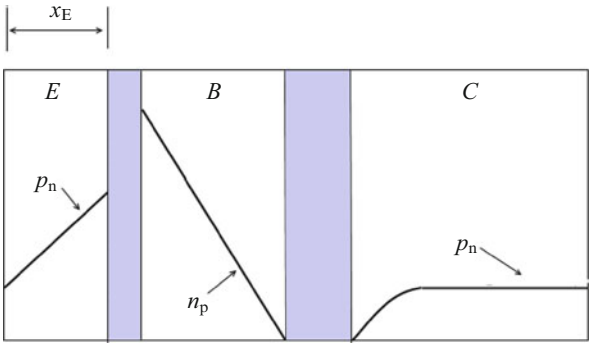


Fig. 3.8 *npn* transistor minority carrier distributions under forward-active bias for the case of a short emitter

is just a forward-biased diode. For a given transistor the length of the emitter region will determine whether to use the long- or short-base diode equations: Referring to Fig. 3.3, if x_E is much greater than the diffusion length for holes, virtually all the injected holes will recombine before reaching the ohmic contact. In this case we found that the excess holes are distributed in the emitter according to an exponentially decaying function (Fig. 3.7) and the resulting hole current is given by

$$I_{pE} = \frac{-qA_E n_i^2 D_p}{N_{dE} L_p} \left(e^{qV_{BE}/k_B T} - 1 \right) \tag{3.20}$$

If on the other hand the emitter contact is close to the base ($x_E << L_p$), the hole concentration is a linear function of x (Fig. 3.8), with a hole current given by the short-base diode result,

$$I_{pE} = \frac{-qA_E n_i^2 D_p}{N_{dE} x_E} \left(e^{qV_{BE}/k_B T} - 1 \right) \tag{3.21}$$

The effectiveness of an emitter junction in injecting electrons into the base is measured by the *emitter injection efficiency*:

$$\gamma = \frac{|I_{nE}|}{|I_{nE}| + |I_{pE}|} = \frac{1}{1 + |I_{pE}/I_{nE}|} \tag{3.22}$$

Since $(|I_{nE}| + |I_{pE}|)$ is the total emitter current I_E, the electron current crossing the emitter–base junction I_{nE} is just γI_E. For a transistor with uniform doping levels and a short emitter we get an emitter injection efficiency of

$$\gamma = \frac{1}{1 + \dfrac{x_B N_{aB} D_{pE}}{x_E N_{dE} D_{nB}}} \tag{3.23a}$$[11]

The above expression can also be used more generally by employing average values for the diffusion coefficients and introducing the Gummel numbers for both the base and the emitter (still under the assumption of a short emitter):

$$\gamma = \frac{1}{1 + \dfrac{G N_B \widetilde{D}_{pE}}{G N_E D_{nB}}} \tag{3.23b}$$

The above results allow us to express the magnitude of the ratio of the collector current I_C to the emitter current I_E under forward-active bias as (neglecting collector leakage current)

$$\alpha_F = \left|\frac{I_C}{I_E}\right| = \frac{\alpha_T |I_{nE}|}{|I_{nE}| + |I_{pE}|} = \alpha_T \gamma \tag{3.24}$$

i.e., the product of the emitter injection efficiency and the base transport factor, which is also known as the *dc alpha* (α_{dc}) of the bipolar transistor. Since all currents into the transistor must sum to zero by Kirchhoff's current law, we have

$$\begin{aligned} I_B + I_E + I_C &= 0 \\ I_B - \frac{I_C}{\alpha_F} + I_C &= 0 \end{aligned} \tag{3.25}$$

or

$$I_C = \frac{\alpha_F I_B}{(1 - \alpha_F)} = \beta_F I_B \tag{3.26}$$

where $\beta_F \equiv I_C/I_B$ is the *dc current gain*, often simply called the beta of the transistor (also commonly denoted by β_{dc} or h_{FE}). Since α_F is near unity for a good transistor, β_F is typically quite large—on the order of 100 or greater.

[11] For a long emitter x_E is replaced by L_p.

Example 3.1: Bipolar Transistor Current Gain Calculation Determine the dc current gain for an idealized silicon *npn* bipolar transistor structure given the following data: $N_{dE} = 2 \times 10^{17}\text{cm}^{-3}$, $N_{aB} = 5 \times 10^{16}\text{cm}^{-3}$, and $\tau_n = 10^{-6}$ s in the base and $\tau_p = 5 \times 10^{-7}$ s in the emitter. The emitter width is 5 μm and the base width is 500 nm.

The base transport factor is given by

$$\alpha_{\text{T}} = 1 - \frac{x_B^2}{2L_n^2} = 1 - \frac{\left(500 \times 10^{-7}\text{cm}\right)^2}{2 \cdot 22.5\ \text{cm}^2\text{s}^{-1} \cdot 10^{-6}\text{s}} = 0.9999444$$

The diffusion length for holes in the emitter is $L_p = \sqrt{D_p \tau_p} = \sqrt{6.5\ \text{cm}^2\text{s}^{-1} \cdot 5 \times 10^{-7}\text{s}} \approx 18\ \mu m$. Therefore the emitter is *short* compared to the diffusion length and the resulting emitter injection efficiency is given by

$$\gamma = \frac{1}{1 + \dfrac{x_B N_{aB} D_{pE}}{x_E N_{dE} D_{nB}}} = \frac{1}{1 + \dfrac{500 \times 10^{-7}\ \text{cm} \cdot 5 \times 10^{16}\text{cm}^{-3} \cdot 6.5\ \text{cm}^2\text{s}^{-1}}{5 \times 10^{-4}\text{cm} \cdot 2 \times 10^{17}\text{cm}^{-3} \cdot 22.5\ \text{cm}^2\text{s}^{-1}}}$$
$$= 0.99283$$

and $\alpha_{\text{F}} = \gamma \alpha_{\text{T}} = 0.99277$.

The dc current gain is therefore $\beta_F = \dfrac{\alpha_{\text{F}}}{1 - \alpha_{\text{F}}} \approx 137$

3.2.2 Operation Modes of a Bipolar Transistor

Thus far we have emphasized that injection of electrons (base minority carriers) into the base region allows current to flow between the collector and emitter (i.e., forward-active bias). To understand bipolar transistor switching we need to consider the distribution and transport of electrons in the base for the different biasing configurations or regions of operation. Based on our analysis of the prototype bipolar transistor structure we can write the general form of the three BJT terminal currents in terms of the applied bias on the two junctions as[12]

$$I_{\text{E}} = \frac{qAn_{\text{i}}^2 D_{\text{nB}}}{x_{\text{B}} N_{\text{aB}}} \left[\exp\left(\frac{qV_{\text{BC}}}{k_{\text{B}}T}\right) - \exp\left(\frac{qV_{\text{BE}}}{k_{\text{B}}T}\right)\right]$$
$$- \frac{qAn_{\text{i}}^2 D_{\text{pE}}}{x_{\text{E}} N_{\text{dE}}} \left[\exp\left(\frac{qV_{\text{BE}}}{k_{\text{B}}T}\right) - 1\right] \tag{3.27a}$$

[12] Equations (3.27) assume a short emitter and collector.

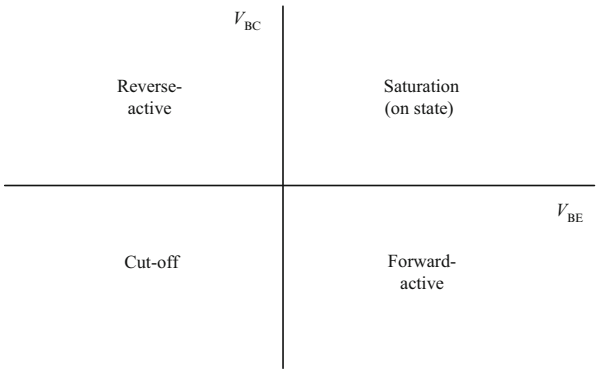


Fig. 3.9 Different quadrants of operation for the biasing configurations of *npn* transistor

$$I_C = \frac{qAn_i^2 D_{nB}}{x_B N_{aB}} \left[\exp\left(\frac{qV_{BE}}{k_B T}\right) - \exp\left(\frac{qV_{BC}}{k_B T}\right)\right] - \frac{qAn_i^2 D_{pC}}{x_C N_{dC}} \left[\exp\left(\frac{qV_{BC}}{k_B T}\right) - 1\right] \tag{3.27b}$$

$$I_B = \frac{qAn_i^2 D_{pE}}{x_E N_{dE}} \left[\exp\left(\frac{qV_{BE}}{k_B T}\right) - 1\right] + \frac{qAn_i^2 D_{pC}}{x_C N_{dC}} \left[\exp\left(\frac{qV_{BC}}{k_B T}\right) - 1\right] \tag{3.27c}$$

The possible bias combinations for an *npn* transistor are illustrated by the different quadrants in Fig. 3.9:

3.2.2.1 Reverse-Active Bias (Quadrant 2)

In the second quadrant, the polarities of V_{BC} and V_{BE} are reversed from those in forward-active bias (quadrant 4). In this region an *npn* transistor injects electrons at the collector and collects them at the emitter, i.e., the roles of the emitter and collector are reversed as illustrated in Fig. 3.10a. A one-to-one correspondence can thus be made to the parameters for forward-active bias. For example, output current is delivered to the emitter lead and the ratio of output to input (base) current in the reverse-active condition is defined as

$$\beta_R = I_E / I_B \tag{3.28}$$

If the transistor is symmetric (like the idealized structure we have been considering) then there is no difference between reverse-active and forward-active mode. However, most practical transistors such as those in integrated circuits are asymmetric in both geometry and doping and they generally perform poorly in reverse-active mode (e.g., they have lower gain).

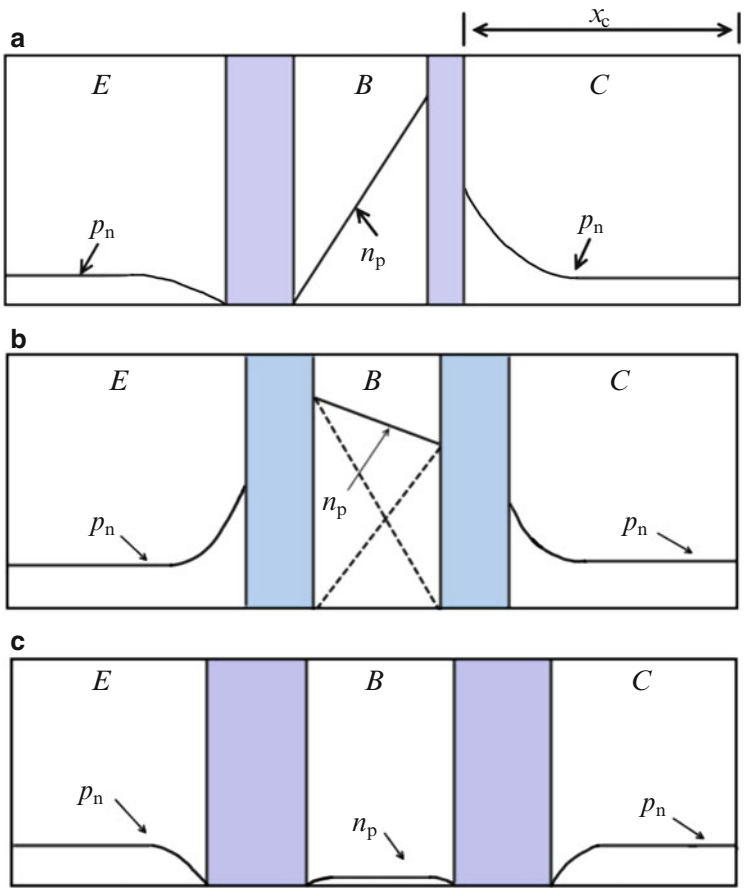


Fig. 3.10 (**a**) *npn* transistor reverse-active bias minority carrier distributions (long collector). (**b**) *npn* transistor minority carrier distributions in saturation. The *dotted lines* indicate that the electron concentration in the base is a superposition of the injected carriers from the forward bias on both junctions. (**c**) *npn* transistor minority carrier distributions in cutoff. The concentration of electrons in the base is very small since it is narrow and will be depleted of minority carriers within a diffusion length of the space-charge regions

3.2.2.2 Saturation (Quadrant 1)

In the first quadrant, both the base–emitter and base–collector junctions are forward biased. This bias condition is called saturation because the current flow is determined by conditions in the circuit external to the transistor, rather than by the transistor itself. In terms of switching, saturation corresponds to the "ON" state of a transistor. We can see from Fig. 3.10b that the saturation condition corresponds to the superposition of forward-active and reverse-active operation. Physically, both junctions are injecting and collecting electrons *at the same time* and electrons are

therefore plentiful throughout the transistor when it is in saturation (essentially like a conducting wire).

3.2.2.3 Cutoff (Quadrant 3)

When V_{BE} and V_{BC} are both negative, the transistor is said to be in the cutoff state. In this condition, both junctions are reverse biased and the base is depleted below its equilibrium population of electrons (Fig. 3.10c). As a result, only very small currents can flow between collector and emitter and the dc behavior of the transistor in cutoff is very close to that of an open circuit.

A transistor switch is biased alternately between regions 1 and 3 (saturation and cutoff) and moves through regions 2 and 4 only during switching transients. To move from saturation to cutoff, the charges stored in and near the base of a transistor must be altered and the two types of *pn* junction capacitance come into play: In cutoff the junction capacitance dominates, whereas in saturation the charge storage or diffusion capacitance will determine the switching speed as we will see in more detail in Sect. 3.6.

3.2.3 Bipolar Transistor I–V Characteristics

The current–voltage characteristics of the BJT depend on how it is connected: Fig. 3.11a shows the *common-base* characteristic where the emitter current is the input and the collector current the output. Figure 3.11b shows the *common–emitter* characteristic where now the base current is the input with collector current the output (this configuration is most often used in applications such as amplification).

Note that I_{CB0} is the collector current flowing with the emitter open circuited ($I_E = 0$). This corresponds to the reverse saturation current of the collector–base junction. The common-base, (forward) active region, collector current can therefore be written

$$I_C = -\alpha_F I_E + I_{CB0} \tag{3.29}$$[13]

Similarly, I_{CE0} is the collector current when the base is open circuited and the common-emitter, (forward) active region, collector current can be written

$$I_C = \beta_F I_B + I_{CE0} \tag{3.30}$$

The collector current can be rewritten in terms of I_{CB0} by substituting for I_E in terms of I_B and I_C in Eq. (3.29):

[13] Recall the emitter current is negative or opposite in sign to the collector current under forward-active bias.

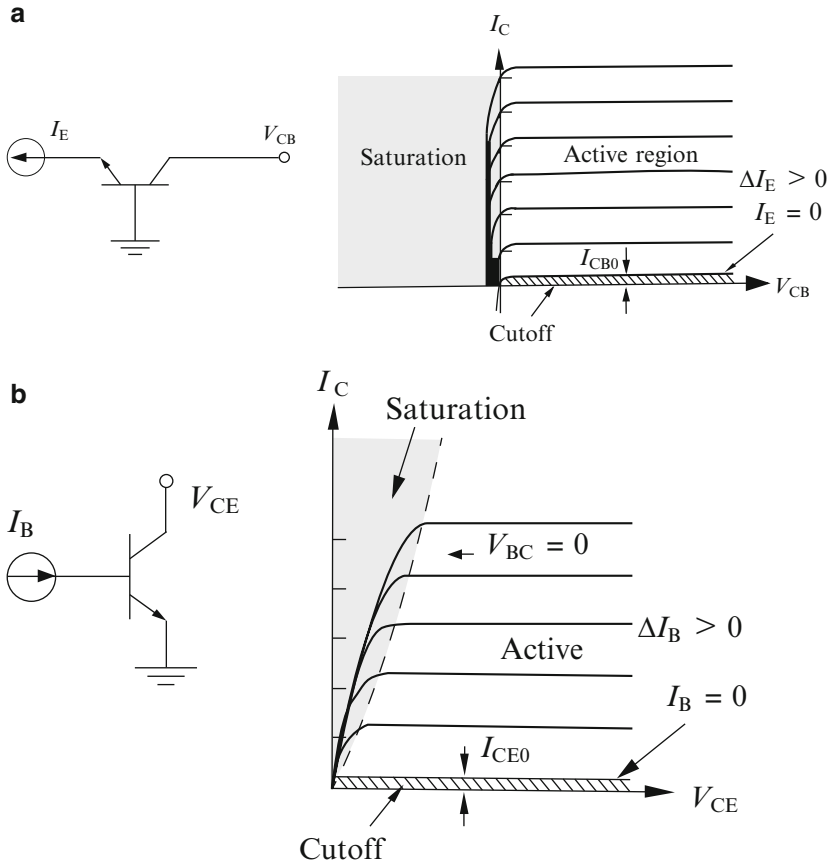


Fig. 3.11 (**a**) *npn* bipolar transistor circuit symbol and common-base characteristic. (**b**) *npn* transistor circuit symbol and common-emitter characteristic

$$\begin{aligned} I_C &= \alpha_F(I_B + I_C) + I_{CB0} \\ \Rightarrow I_C &= \frac{\alpha_F}{1-\alpha_F} I_B + \frac{I_{CB0}}{1-\alpha_F} \end{aligned} \tag{3.31}$$

and therefore,

$$I_{CE0} = \frac{I_{CB0}}{1-\alpha_F} \tag{3.32}$$

We thus see that the collector current that flows with an open base is roughly a factor of beta larger than the corresponding current that flows when the emitter is open. We will return to this point later in the next section.

3.3 Ebers–Moll Model

In order to describe the basic properties of a bipolar transistor and the shape of its *I–V* characteristic a useful model was developed by Ebers and Moll in 1954. This model still provides the essential framework for computer-aided simulation of bipolar transistors with arbitrary bias configurations and in large circuits.

The Ebers–Moll model is based on treating the BJT as two interacting or coupled *pn* diodes in order to extract a set of parameters that describe the bipolar transistor: Previously, we wrote the electron current flowing between the emitter and collector as

$$I_{\mathrm{nE}} = I_0\left[\exp\left(\frac{qV_{\mathrm{BC}}}{k_{\mathrm{E}}T}\right) - \exp\left(\frac{qV_{\mathrm{BE}}}{k_{\mathrm{B}}T}\right)\right] \tag{3.33}$$

This current connects the emitter and collector and is sometimes referred to as the *linking current* (in a *pnp* transistor this would instead be I_{pE}). The current flowing between the base and emitter can be written

$$I_{\mathrm{BE}} = I_{0\mathrm{E}}[\exp(qV_{\mathrm{BE}}/k_{\mathrm{B}}T) - 1] \tag{3.34}$$

since the base–emitter junction is a *pn* diode. The total current in the emitter consists of the flow to the collector (the linking current) minus the base–emitter diode current:

$$I_{\mathrm{E}} = I_0[\exp(qV_{\mathrm{BC}}/k_{\mathrm{B}}T) - \exp(qV_{\mathrm{BE}}/k_{\mathrm{B}}T)] - I_{0\mathrm{E}}[\exp(qV_{\mathrm{BE}}/k_{\mathrm{B}}T) - 1] \tag{3.35}$$

Similarly, for the collector current we have

$$I_{\mathrm{C}} = I_0[\exp(qV_{\mathrm{BE}}/k_{\mathrm{B}}T) - \exp(qV_{\mathrm{BC}}/k_{\mathrm{B}}T)] - I_{0\mathrm{C}}[\exp(qV_{\mathrm{BC}}/k_{\mathrm{B}}T) - 1] \tag{3.36}$$

where the base–collector current is

$$I_{\mathrm{BC}} = I_{0\mathrm{C}}[\exp(qV_{\mathrm{BC}}/k_{\mathrm{B}}T) - 1] \tag{3.37}$$

Grouping terms in I_E and I_C gives

$$\begin{aligned} I_{\mathrm{E}} &= -(I_0 + I_{0\mathrm{E}})[\exp(qV_{\mathrm{BE}}/k_{\mathrm{B}}T) - 1] + I_0[\exp(qV_{\mathrm{BC}}/k_{\mathrm{B}}T) - 1] \\ I_{\mathrm{C}} &= -(I_0 + I_{0\mathrm{C}})[\exp(qV_{\mathrm{BC}}/k_{\mathrm{B}}T) - 1] + I_0[\exp(qV_{\mathrm{BE}}/k_{\mathrm{B}}T) - 1] \end{aligned} \tag{3.38}$$

We can now define

$$I_{F0} \equiv I_0 + I_{0E},\ I_{R0} \equiv I_0 + I_{0C} \tag{3.39a}$$

and

$$\alpha_F \equiv \frac{I_0}{I_0 + I_{0E}},\ \alpha_R \equiv \frac{I_0}{I_0 + I_{0C}} \tag{3.39b}$$

In terms of these new variables the current equations become

$$I_E = -I_{F0}[\exp(qV_{BE}/k_BT) - 1] + \alpha_R I_{R0}[\exp(qV_{BC}/k_BT) - 1] \tag{3.39c}$$

$$I_C = -I_{R0}[\exp(qV_{BC}/k_BT) - 1] + \alpha_F I_{F0}[\exp(qV_{BE}/k_BT) - 1] \tag{3.39d}$$

Equations (3.39) are the Ebers–Moll equations for an *npn* transistor. (In the corresponding equations for a *pnp* transistor the current directions are reversed and V_{BE} and V_{BC} are changed to V_{EB} and V_{CB}, respectively.)

The Ebers–Moll equations directly predict the emitter and collector currents for the transistor and in conjunction with Kirchhoff's current law they also specify the base current. The model has four parameters; however, only three are independent since

$$\alpha_F I_{F0} \equiv \alpha_R I_{R0} \equiv I_0 \tag{3.40}$$

which is known as the *reciprocity relation*. The Ebers–Moll equations can be written in more compact form by defining two new quantities:

$$I_F = I_{F0}[\exp(qV_{BE}/k_BT) - 1] \tag{3.41a}$$

as a diode current related to forward-active bias, and

$$I_R = I_{R0}[\exp(qV_{BC}/k_BT) - 1] \tag{3.41b}$$

related to reverse-active bias. In terms of I_F and I_R the Ebers–Moll equations take the form

$$I_E = -I_F + \alpha_R I_R \tag{3.41c}$$

$$I_C = -I_R + \alpha_F I_F \tag{3.41d}$$

These equations can be represented by the equivalent circuit shown in Fig. 3.12 consisting of diodes and current sources connected between the base and the emitter and the base and the collector, with α_F and α_R coupling the emitter and collector to each other. Applying Kirchhoff's current law to this circuit allows us to solve for the base current:

$$I_B = -(I_E + I_C) = I_F(1 - \alpha_F) + I_R(1 - \alpha_R) \tag{3.41e}$$

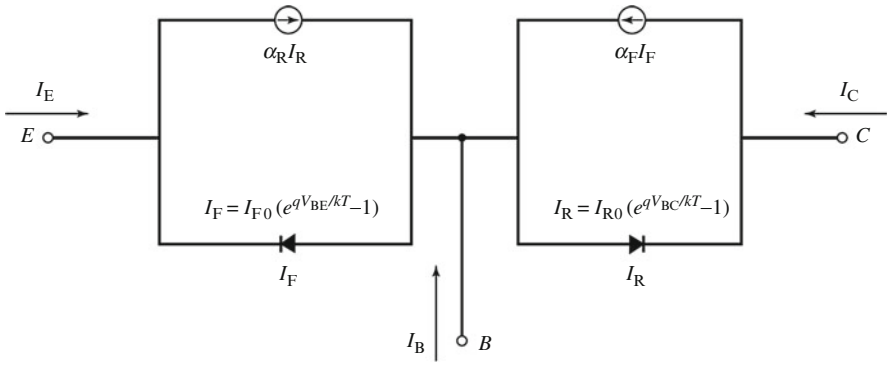


Fig. 3.12 Ebers–Moll equivalent circuit for an *npn* bipolar transistor

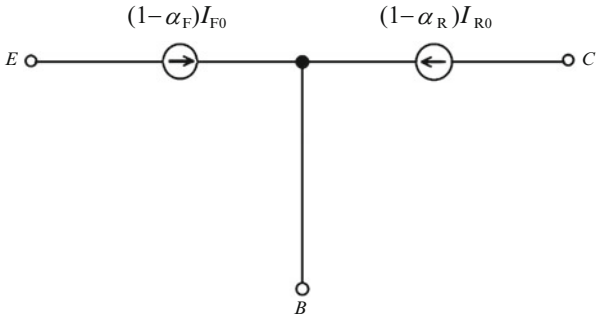


Fig. 3.13 Ebers–Moll equivalent circuit for an *npn* transistor in cut-off (The reciprocity relation (Eq. 3.40) was used to derive this result)

Example 3.2: Ebers–Moll Equations To see how the Ebers–Moll model represents the transistor in various regions of operation, we can consider some examples:

1. Cutoff
   In this region V_{BE} and V_{BC} are both negative and from the Ebers–Moll equations we obtain the equivalent circuit shown in Fig. 3.13. That is, the model reduces to two current sources that represent the reverse saturation currents of the two junctions.
2. Active
   We can rearrange the equations to express the collector current in terms of the emitter current:

$$I_C = -\alpha_F I_E - I_R(1 - \alpha_F \alpha_R) \tag{3.43a}$$

   which under forward active-bias becomes

$$I_C = -\alpha_F I_E + I_{R0}(1 - \alpha_F \alpha_R) \tag{3.43b}$$

Similarly, reverse-active bias results in

$$I_E = -\alpha_R I_C + I_{F0}(1 - \alpha_F \alpha_R) \tag{3.44}$$

These equations show that the four parameters of the Ebers–Moll model can be obtained by measuring I_C versus I_E under forward-active bias and I_E versus I_C under reverse-active bias. If forward-active measurements are used, a plot of I_C versus I_E should be linear with slope equal in magnitude to α_F. The intercept with $I_E = 0$ corresponds to a measurement with open-circuited emitter, and the current flowing in this case is given by

$$I_{CB0} = I_C\big|_{I_E=0} = I_{R0}(1 - \alpha_F \alpha_R) \tag{3.45}$$

We can compare this to the current I_{CE0} in the collector when the base is open circuited by substituting $I_E = -I_C$ in the equation for collector current:

$$I_{CE0} = I_C\big|_{I_B=0} = \frac{I_{R0}(1 - \alpha_F \alpha_R)}{(1 - \alpha_F)} = \frac{I_{CB0}}{(1 - \alpha_F)} \tag{3.46}$$

as we found in Eq. (3.32) above. The reason for the difference in magnitudes between I_{CB0} and I_{CE0} is related to the boundary conditions determined by the bias on the emitter–base junction: When I_{CB0} flows, the emitter–base junction becomes slightly reverse biased because some electrons are extracted from the emitter without being replaced from the external circuit (see Problem 3). In this case the collector current is only the reverse leakage current. When I_{CE0} flows, the base–emitter junction instead becomes forward biased, and most of the collector current results from electrons carried across the base from the emitter. The leakage current I_{CB0} is therefore effectively multiplied by the transistor gain.

Lastly, by using the Ebers–Moll equations equivalent circuits for active bias in the common-base and common-emitter configuration of an *npn* transistor can be found as depicted in Fig. 3.14.

3. Saturation

In saturation the quantity of greatest importance is V_{CEsat}, the voltage drop across the "ON state" switch. In this case the Ebers–Moll model allows one to derive

$$V_{CEsat} = \frac{k_B T}{q} \ln\left\{ \frac{\left[1 + \frac{I_C}{I_B}(1 - \alpha_R)\right]}{\alpha_R \left[1 - \frac{I_C}{I_B}\left(\frac{1-\alpha_F}{\alpha_F}\right)\right]} \right\} \tag{3.47}$$

This equation shows that V_{CEsat} is a weak function of the collector current. Note that the Ebers–Moll equations do not take into account any resistance in series with the junctions. The voltage drops across these resistances, especially in the collector region, often exceed the value of V_{CEsat} predicted by the above equation. Thus the "ON" state of a transistor switch is often modeled by including a series resistor denoted R_{Csat}.

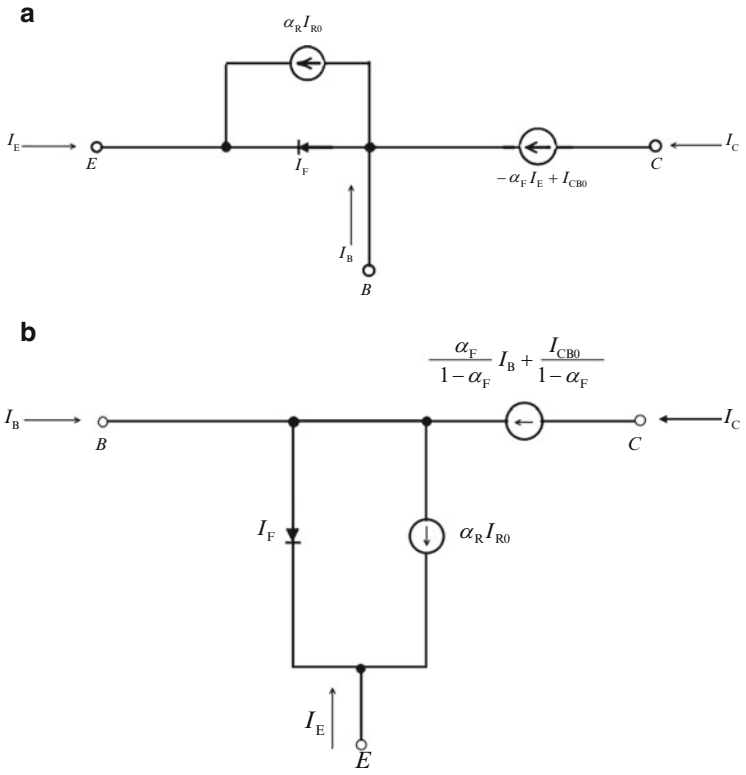


Fig. 3.14 (**a**) Ebers–Moll equivalent circuit for an *npn* bipolar transistor in the common-base configuration. (**b**) Ebers–Moll equivalent circuit for an *npn* bipolar transistor in the common-emitter configuration

3.4 Deviations from Ideal Behavior

3.4.1 Base-Width Modulation

In real transistors under forward-active bias the voltage across the collector–base junction influences collector current, primarily through what is known as the *Early effect*[14]: Previously, when we considered bipolar transistors under active bias the voltage applied across the collector–base junction was simply to ensure the efficient collection of base minority carriers and their delivery to the collector region. However, we know from the results in Chap. 2 on *pn* junctions that the width of the space-charge region is voltage dependent. Thus, in the case of bipolar

[14] J. M Early, Proc. IRE **40**, 1401 (1952).

transistors, a changing collector–base bias varies the depletion layer width at the collector junction and consequently the width of the neutral base region. This variation complicates the performance of the transistor as a linear amplifier since from our previous results (Eqn. 3.11) under forward-active bias we have

$$I_C = \frac{q\widetilde{D}_n n_i^2 A_E \exp(qV_{BE}/k_B T)}{\int_0^{x_B} p dx} \tag{3.48}$$

where the integration is performed over x_B, the width of the neutral base region. Thus the Early effect results in an increase in I_C due to base-width modulation when the magnitude of reverse bias on the collector V_{CB} is increased, which is evident if we examine the common-emitter output characteristic of the BJT shown in Fig. 3.15a. Physically, the Early effect can be understood by considering the transistor structure in active bias as illustrated in Fig. 3.15b: We have seen that the minority carrier distribution in the base will have a triangular shape. Increasing the collector–base voltage moves the base–collector junction depletion layer edge towards the emitter and a second triangular distribution specifies the new minority carrier profile; thus the gradient of n_p is increased, which in turn increases the collector current. We can also see that the Early effect reduces charge storage in the base and therefore in addition to the change in collector current, a decrease in the base minority charge results in a smaller base current since recombination is reduced [see Eq. (3.16)].

3.4.2 Punchthrough

We have seen that reducing the base width and decreasing the base doping both increase gain. However, as the base width is reduced to submicron dimensions the applied base–collector voltage can deplete the *entire* width of a moderately doped neutral base region so that the collector–base space-charge region reaches through to the *emitter–base* space-charge region at higher voltages. This will reduce the barrier at the emitter–base junction resulting in a highly conductive path for carriers from emitter to collector under active bias as illustrated in Fig. 3.16 (many carriers are available in the emitter to participate in current flow). This effect is called *punchthrough* and can lead to damaging currents in a transistor and should generally be avoided by proper transistor design (see Problem 4).

3.4.3 Reverse-Bias Breakdown

Another failure mode that is sensitive to V_{CB} is avalanche breakdown of the collector, similar to what we saw earlier in connection with *pn* diodes. In a bipolar transistor, however, the breakdown voltage can be significantly lower than that of the analogous

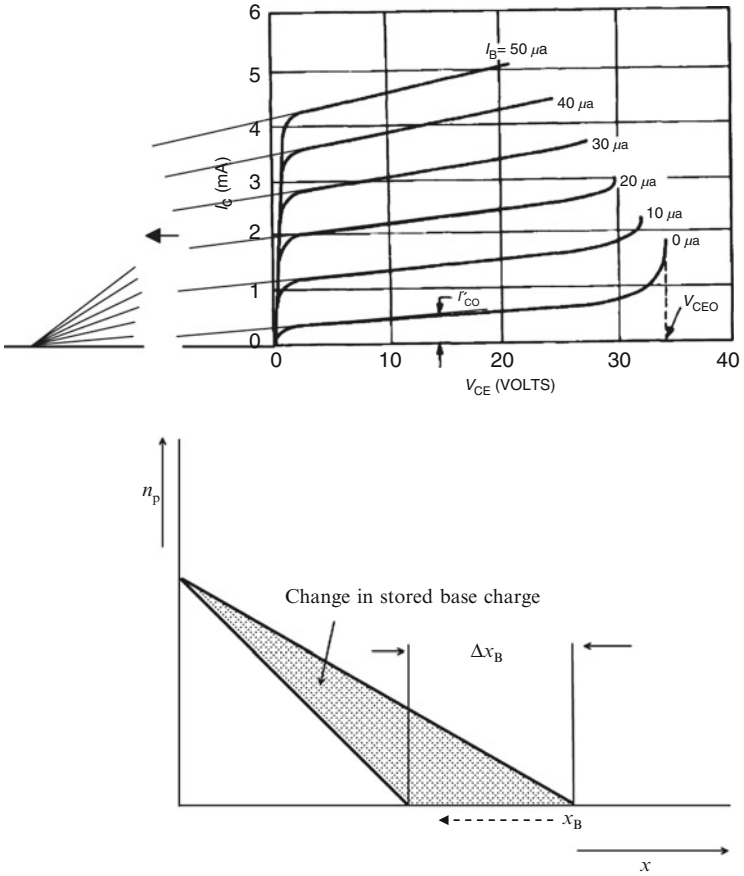


Fig. 3.15 (**a**) Early effect displayed as an increase in collector current for the common-emitter *I–V* characteristic. The Early voltage, V_A, is found by extrapolating the constant slope regions of the curves as illustrated (Adapted from M. J. Morant, *Introduction to Semiconductor Devices*, Addison-Wesley, Reading, 1964.) (Note that $V_{CE} \approx V_{CB} + 0.7$ V for bias in the forward-active region of a silicon transistor). (**b**) Illustration of the variation in minority carrier distribution within the base of a bipolar transistor due to base-width modulation, which leads to the Early effect and also reduces the amount of charge stored in the base

reverse-biased base–collector diode. The reason for this can be understood by considering an *npn* transistor under active bias with the base lead open: We saw earlier [Eqs. (3.32) and (3.46)] that in this case the emitter–base junction became forward biased leading to a current:

$$I_{CE0} = \frac{I_{CB0}}{(1 - \alpha_F)}$$

Electrons entering the base–collector junction are accelerated by the high electric field in this reverse-biased junction and can create new electron–hole pairs via the

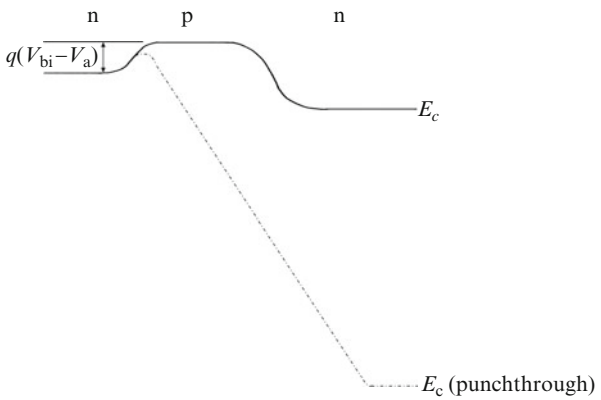


Fig. 3.16 *npn* bipolar transistor conduction band edge for the forward-active bias punchthrough breakdown condition (band edge slopes not to scale)

impact ionization process described in Chap. 2. According to the above equation, the additional current due to these new carriers is then multiplied by the gain, and this process repeats itself, etc., leading to an avalanche breakdown that occurs at a much lower voltage than the base–collector diode breakdown voltage itself. Qualitatively, the lower breakdown voltage occurs because of positive feedback from the bipolar gain.

The breakdown voltage is usually denoted by BV_{CE0} (the collector–emitter breakdown voltage with the base lead open) and can be related to the base–collector diode breakdown voltage (BV_{CB0}) by the expression:

$$BV_{CE0} = \frac{BV_{CB0}}{\beta^{1/m}} \tag{3.49}$$

Empirically, m is typically found to be about 4 and some typical data is shown in Fig. 3.17.

3.4.4 *Effects at Low and High Emitter Bias*

The theory we have developed for bipolar transistors thus far predicts that the collector and base currents depend exponentially on base–emitter voltage, i.e.,

$$I_C \approx I_0 \exp\left(\frac{V_{BE}}{V_t}\right); V_t \equiv k_B T/q \tag{3.50}$$[15]

and similarly for base current with just a different (smaller) factor multiplying the exponent. If we plot experimental data for an *npn* transistor in the forward-active region on a semi-log scale there is an excellent fit to straight lines over the

[15] V_t is usually referred to as the "thermal voltage."

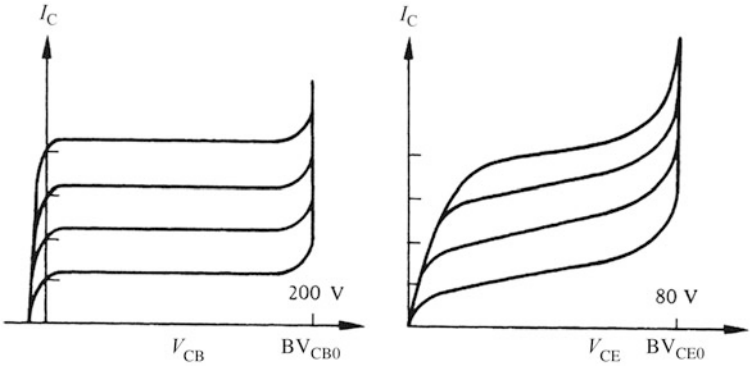


Fig. 3.17 Bipolar transistor breakdown voltage comparison for the common-base (left) and common-emitter (right) configurations (Note, in addition, the absence of an appreciable Early effect in the common-base characteristic) (Adapted from G.W. Neudeck, The Bipolar Junction Transistor, 2nd edn. Prentice-Hall, Boston, 1989)

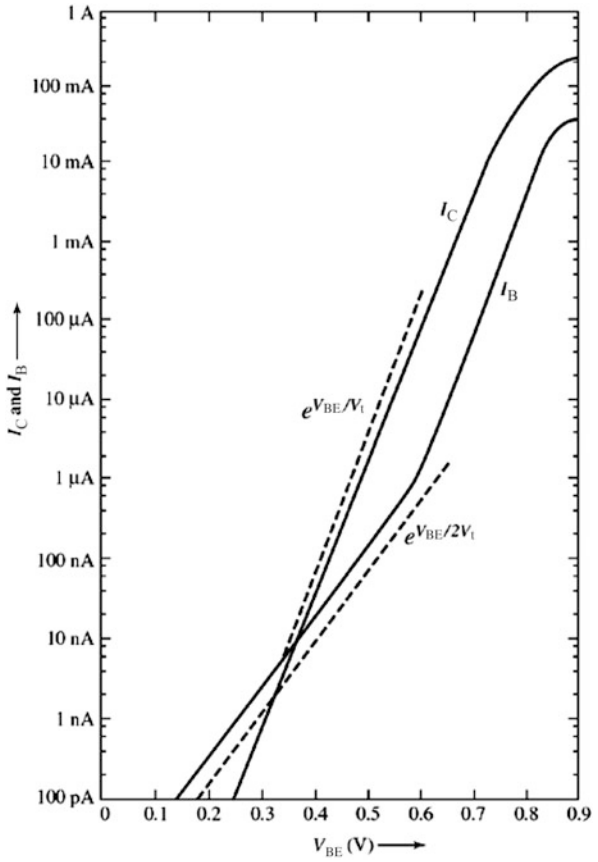


Fig. 3.18 Collector and base currents as a function of base–emitter voltage (After A. S. Grove, *Physics and Technology of Semiconductor Devices*, Wiley, 1967)

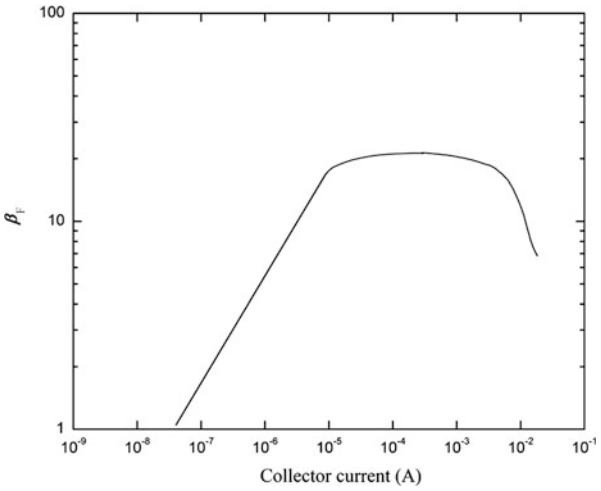


Fig. 3.19 Illustration of current gain vs. collector current for a bipolar transistor

midrange of currents, consistent with the equation above. Straight lines of constant slope do not however adequately represent all of the data for either I_C or I_B as illustrated in Fig. 3.18.

3.4.4.1 Base–Emitter Space-Charge Region Recombination

We can see that the collector current is "ideal" over a wider voltage range than the base current. The variation in I_B at low voltages is related to recombination in the base–emitter space-charge region, which, as we already saw for *pn* junctions, causes excess current compared to ideal diode theory at low biases. Thus for V_{BE} less than about 0.5 V we can write

$$I_B \propto \exp\left(\frac{qV_{BE}}{\eta k_B T}\right) \tag{3.51}$$

with η approaching 2 as voltage is decreased. Recombination current in the space-charge region flows only in the base and emitter leads—it does not affect the collector current, which consists almost entirely of electrons that are injected across the base from the emitter junction. Thus the collector current will continue to be represented by the ideal exponential dependence as V_{BE} decreases. The bipolar transistor effect is essentially able to "separate out" this non-ideality from the output collector current. (However, at some point injection will become so low that the collector current is dominated by thermal carrier generation in the *collector* space-charge region since it is reverse biased.)

Thus, at low biases the collector current is a smaller fraction of emitter current than in the intermediate bias range. This behavior can be seen more clearly in Fig. 3.19 where

$$|I_C/I_B| \equiv \beta_F$$

is plotted using the data from Fig. 3.18. The decrease in β_F as base–emitter bias decreases is a limitation on the use of bipolar transistors to amplify low voltages. Maintaining large current gains at low bias levels (and thus reducing space-charge recombination current) is important in many applications. For example, transistors and integrated circuits used in hearing-aid amplifiers, pacemakers, and other therapeutic or diagnostic biomedical systems depend on achieving adequate performance at the lowest possible currents (hence minimizing power dissipation). For these and other types of specialized applications material and device processing efforts focus on maintaining as high a lifetime as possible within the space-charge regions.

3.4.4.2 High-Level Injection at the Base–Emitter Junction

The basic equation for collector current under active bias is seen in Fig. 3.18 to be valid over nearly eight decades of current until it deviates at high base–emitter bias. One cause for this deviation is that the assumption of low-level injection used to derive the ideal diode equation is no longer valid at high bias. High-level injection of electrons into the base will cause a significant increase in the population of base majority carriers and therefore the integrated hole charge in the base will depend on applied bias. Thus the collector current,

$$I_C = \frac{q\widetilde{D}_n n_i^2 A_E \exp(qV_{BE}/k_B T)}{\int_0^{x_B} p dx},$$

will decrease for large applied bias, similar to high-level injection in *pn* junctions.[16] Another effect that becomes important at high currents is the *base spreading resistance* or current crowding phenomenon illustrated in Fig. 3.20: The finite resistance of the base causes the base–emitter voltage to vary with position, decreasing away from the base contact, due to the ohmic voltage drop. This causes the collector current to deviate from ideal behavior at high currents similar to a forward-biased *pn* junction, as shown in Fig. 3.21.

We have only considered some of the effects which limit bipolar transistor performance. In practice, these and other effects (thermal effects, base–collector space-charge region modification at high currents,[17] etc.) will occur simultaneously and there can be interactions between them. Computer-aided analysis can treat these effects in a self-consistent manner. The standard way this is done in practice is to modify the Ebers–Moll model to include (1) recombination in the emitter–base space-charge region,

[16] See Eq. (2.45).

[17] For example, if large numbers of electrons are swept through the base–collector junction under forward-active bias there will be a significant average negative charge added to the fixed charges in the depletion layer. Qualitatively, this results in a net charge increase in the negative side of the depletion layer and a net decrease in the positive side, which means the depletion width on the p-side ($-x_p$) becomes *smaller* and leads to an *increase* in the base width, x_B, at large collector currents, known as the *Kirk effect* (C. T. Kirk, IRE Trans. Electron Devices **ED-9**, 164 (1962).

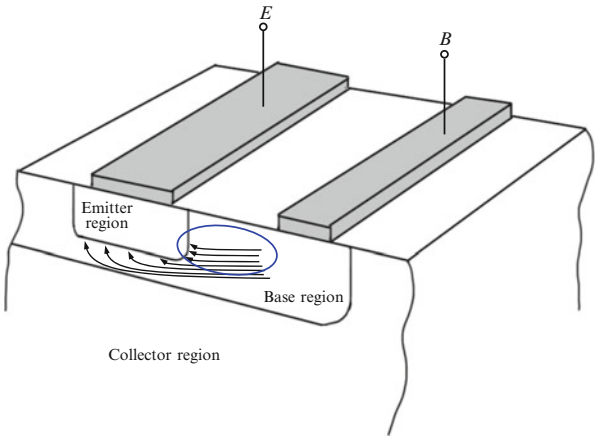


Fig. 3.20 Three-dimensional schematic cross-section of a bipolar transistor illustrating base spreading resistance. Current crowding occurs near the edge of the base contact (encircled region) as majority carriers take the path of least resistance to the emitter (collector contact not shown)

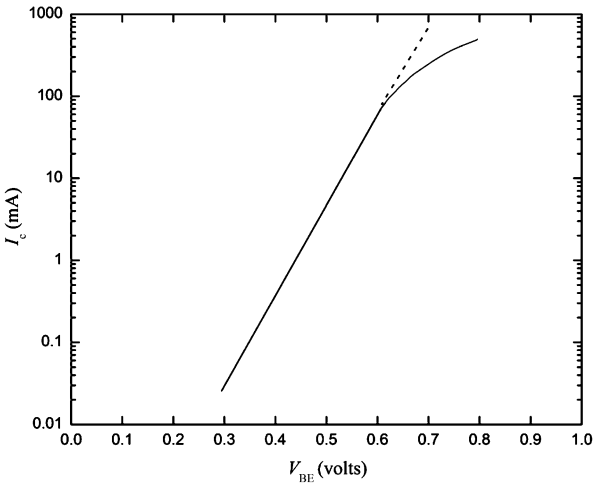


Fig. 3.21 Collector current decrease at large emitter bias

(2) current-gain decrease under high-current conditions, and (3) effects of space-charge layer widening (Early effect). This modified approach including various second-order effects is known as the *Gummel–Poon model*.

3.5 Small-Signal Parameters[18]

Thus far the bipolar transistor models we have considered have been static ones, i.e., for dc conditions. To allow for dynamical calculations we start by examining the small-signal behavior of the BJT, similar to what we did for the *pn* diode in Chap. 2.

[18] Similar to the small-signal treatment of Chap. 2, the results of this section are based on ideal step *pn* junctions unless otherwise noted. Non-idealities related to the *pn* junctions themselves can also be included if necessary.

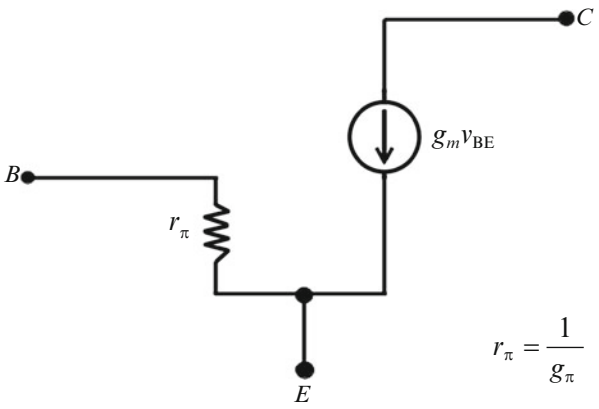


Fig. 3.22 Low-frequency first-order BJT equivalent circuit

When bipolar transistors are biased in the active region and used for amplification it is useful to approximate their behavior under conditions of small voltage variations at the base–emitter junction. The *transconductance* g_m or g_{fe} of the transistor expresses the effectiveness of the control of the collector current by the base–emitter voltage:

$$g_m \equiv \frac{\partial I_C}{\partial V_{BE}} \tag{3.52a}$$

Hence, using Eq. (3.50) the BJT transconductance is given by

$$g_m = \frac{I_0}{V_t} \exp\left(\frac{V_{BE}}{V_t}\right) = \frac{I_C}{V_t} \tag{3.52b}$$

Note that g_m is directly proportional to the collector (or output) current. We can also express the variation of base current with base–emitter voltage as

$$g_\pi \equiv \frac{\partial I_B}{\partial V_{BE}} = \frac{1}{\beta_F} \frac{\partial I_C}{\partial V_{BE}} = \frac{g_m}{\beta_F} \tag{3.53}$$

If we denote the small-signal (ac) voltages by lowercase symbols, a simple low-frequency (i.e., ignoring capacitance) small-signal active bias BJT equivalent circuit thus takes the form shown in Fig. 3.22. This equivalent circuit emphasizes that, to first order, the input (base) current is decoupled from the output (collector) and that the output is insensitive to collector–base voltage variations.

However, in real transistors we have seen that the voltage across the collector–base junction does influence collector current via the Early effect (base-width modulation).

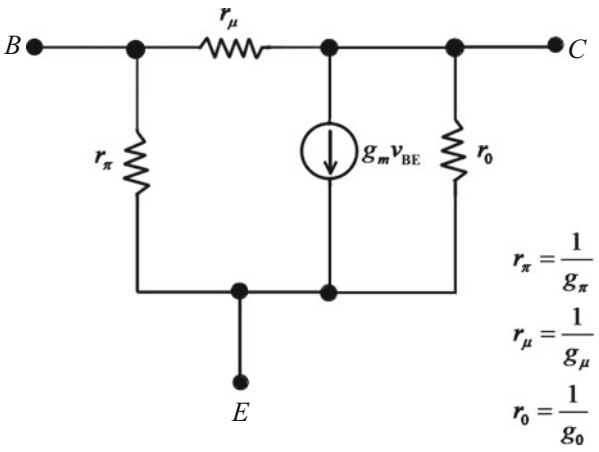


Fig. 3.23 Low-frequency BJT equivalent circuit including the Early effect

The variation of I_C with V_{CB} can be expressed as the ratio of the collector current to the *Early voltage* V_A:

$$g_0 \equiv \left|\frac{\partial I_C}{\partial V_{CB}}\right| \equiv \frac{I_C}{|V_A|} = \frac{g_m V_t}{|V_A|} \tag{3.54a}$$

where Eq. (3.52b) has been used to express the collector current in terms of the transconductance. Similarly, we saw in the previous section that the change in base minority charge also resulted in a change in base current and therefore

$$g_\mu \equiv \left|\frac{\partial I_B}{\partial V_{CB}}\right| = \frac{g_m V_t}{|V_A|\beta_F} \tag{3.54b}$$

These variations due to the Early effect can be incorporated into the low-frequency small-signal equivalent circuit shown in Fig. 3.23.

Since the base–emitter junction is forward biased in the active mode it has a capacitance associated with incremental changes in the injected minority carrier charge, i.e., the diffusion capacitance, C_d ($\equiv C_\pi$), that we saw earlier for a *pn* junction. We have seen that variation in I_C with V_{CB} must also cause a change in the amount of charge stored in the base. This can also be included through an additional capacitance leading to the equivalent circuit in Fig. 3.24, known as the *hybrid-pi circuit model* for the bipolar transistor. A more accurate (and complicated) equivalent circuit representation would also take space-charge junction capacitances and series resistances into account along with other parasitic contributions.

3.5.1 Frequency Limits of Bipolar Transistors

At higher frequencies the gain of a transistor decreases because of the finite time needed to move charges to various parts of the transistor and the equivalent circuit models described above will begin to break down. By considering the movement of

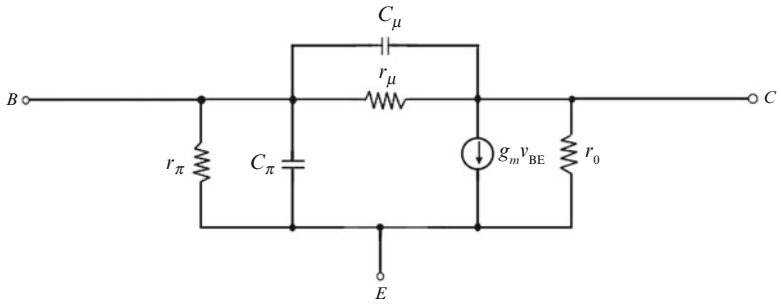


Fig. 3.24 Hybrid-pi BJT equivalent circuit

charges into different regions of the transistor and the rearrangement of these charges we can estimate the frequency dependence of the gain. The associated delay times can be found by dividing the charge that must be moved by the current that moves it.

The total delay time to charge the transistor can be written

$$\tau_{EC} = 1/2\pi f_T \tag{3.55}$$

which can be expressed in terms of characteristic delay times in the collector, base, and emitter regions of the transistor. We consider the most important delay times below.

The *base transit time* for a uniformly doped *npn* transistor follows from the short-base diode result, viz.,

$$\tau_t \equiv \tau_B = \frac{Q_B}{I_C} = \frac{x_B^2}{2D_{nB}} \tag{3.56}$$

In modern transistors the base is very thin so other delay times usually become the limiting factors to the ultimate speed of the transistor. Carriers will also spend a finite length of time traveling through the base–collector depletion region of width x_{dC}. Due to the large electric field present inside the junction the carriers will quickly reach their saturation velocity and thus the *collector transit time* can be written

$$\tau_C \approx \frac{x_{dC}}{v_{sat}} \tag{3.57}$$

Finally, in addition to the above delay times associated with mobile charges, we also add the *RC* delays associated with the two space-charge region junction capacitances, resulting in an expression for the *cutoff frequency* of a BJT:

$$f_T = (2\pi\tau_{EC})^{-1} = [2\pi(\tau_{RC_E} + \tau_C + \tau_B + \tau_{RC_C})]^{-1} \tag{3.58}$$

which corresponds to the frequency at which the gain of the transistor becomes unity.

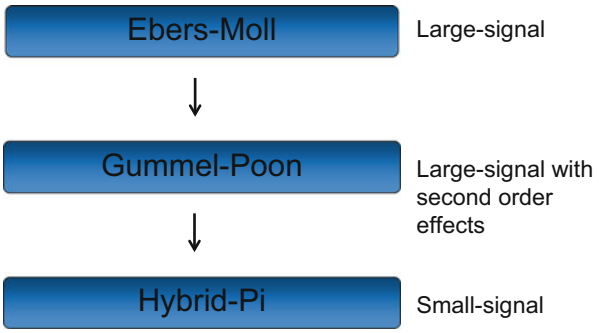


Fig. 3.25 Standard bipolar transistor models

Example 3.3: Bipolar Transistor Cutoff Frequency Calculation Assume the transit time for electrons across the base of a silicon *npn* bipolar transistor is 50 ps. The emitter–base junction charging time is 20 ps and the collector capacitance and resistance are 0.3 pF and 20 Ω, respectively. Assume also that the collector junction depletion region is 1 μm wide. Find the cutoff frequency of the transistor.

$$\frac{1}{2\pi f_T} = \tau_{\mathrm{EC}} = \left[(50+20)\times 10^{-12} + 0.3\times 10^{-12}\cdot 20 + \frac{10^{-4}}{10^{7}}\right]\mathrm{s} = 86\ \ \mathrm{ps}$$

and therefore the cutoff frequency is

$$f_T = \frac{1}{2\pi\tau_{\mathrm{EC}}} \approx 1.85\ \ \mathrm{GHz}$$

Figure 3.25 summarizes the various bipolar transistor models and their applications.

3.6 Optimizing Bipolar Transistor Design and Performance

3.6.1 Performance Versus Device Structure

The table in Fig. 3.26 contains a list of possible bipolar transistor performance requirements and the doping levels that are needed to achieve them. It is immediately evident from this table that all requirements cannot be met simultaneously and trade-offs in the design of a BJT are necessary in practice. Depending on whether bipolar transistors are used for amplification or switching the relative importance of these requirements will change. Because of the significantly higher mobility of

Requirement	E-doping	B-doping	C-doping
High emitter injection efficiency	high	low	
Low base resistance		high	
Low base-width modulation (Early effect)		high	low
Low collector leakage		high	high
High collector breakdown voltage		low	low
Low collector resistance			high

Fig. 3.26 Bipolar transistor performance requirements vs. doping levels

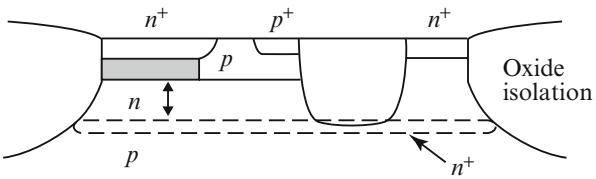


Fig. 3.27 Bipolar transistor IC structure. The gray region denotes the active part of the base. The width and doping of the collector (labeled by the double arrow) determine transistor performance for amplifying or switching applications.

electrons vs. holes in most semiconductors,[19] *npn* transistors are used more frequently than *pnp* transistors.

For integrated circuits, the thickness and resistivity of the collector region as illustrated in Fig. 3.27 will depend on the intended application. For *amplifying BJTs*, the thickness and resistivity of the collector are both larger. This results in an increased breakdown voltage and reduces the Early effect and junction capacitance. For *switching BJTs*, saturation ("ON state") resistance must be minimized, which requires a very thin collector layer with a low resistivity (usually much less than 1 Ω-cm).

[19] This makes sense intuitively since the motion of a hole corresponds to the movement of many electrons in the valence band (see Appendix A) and thus has a larger "inertia" compared to the free electrons in the conduction band.

3.6.2 Transient/Switching Behavior

The large amount of charge stored in the base and collector at saturation can cause considerable delay during switching. The time required for the transistor to switch from the cutoff state to saturation and vice versa (Fig. 3.28a) will depend on how quickly excess minority charge carriers can be added and removed from the transistor (Fig. 3.28b). As in the case of a *pn* junction diode, the minority carrier lifetimes/transit times will be key parameters that determine how fast a transistor can be switched.

One means of speeding the removal of transistor charge is the incorporation of a large number of recombination centers to reduce the carrier lifetime. This can be done for example by doping the collector region with gold impurity atoms. More often, the technique of *Schottky clamping* is instead employed to obtain superior switching results by drastically reducing the amount of excess charge stored in the base and the collector, without requiring additional impurities. The Schottky-clamped transistor is a normal bipolar transistor with a Schottky diode connected between the base and collector (Fig. 3.28c). Because the Schottky diode turn-on voltage is generally lower than a *pn* junction diode (see Chap. 2), most of the excess base current will be shunted through the Schottky diode instead of the base–collector junction, greatly reducing the amount of excess charge in the base and collector. This results in much improved switching times (Fig. 3.28d). In addition, the clamped transistor takes up only slightly more surface area on a silicon chip compared to an unclamped transistor (Fig. 3.29).

3.6.3 Emitter Injection Efficiency

We have seen that the emitter is usually heavily doped to increase the emitter injection efficiency. However, when doped above approximately 10^{19} cm^{-3} efficiency *decreases* due to the phenomenon of *band gap narrowing* illustrated in Fig. 3.30.[20] In addition, we would like to have the base doped as lightly as possible. However if the doping is too low the series resistance will start to become a factor.

These effects can be alleviated by using a wide band gap emitter and a narrow band gap base. This type of device is known as a *heterojunction bipolar transistor* (HBT) (Fig. 3.31a). HBTs can have current gains several orders of magnitude larger than regular bipolar transistors and with proper design are able to operate at several hundred GHz or more. A further refinement that can be combined with the HBT structure is to employ a built-in electric field in the base of the bipolar transistor in

[20] A smaller band gap in the emitter means that the intrinsic carrier concentration will increase, which leads to an increase in the amount of holes injected into the emitter of an *npn* transistor and thus lowers the emitter injection efficiency.

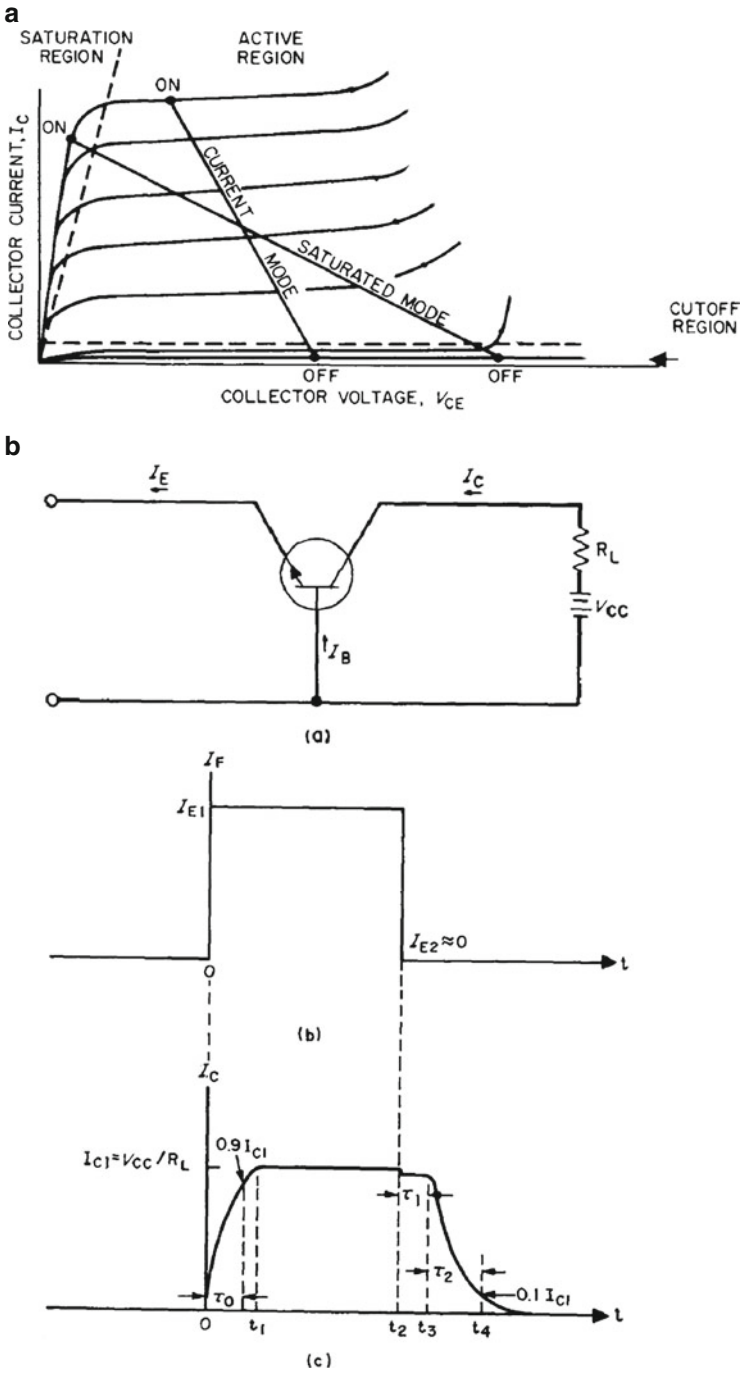


Fig. 3.28 BJT switching behavior. (**a**) Illustration of on and off states based on the load line of the transistor circuit (Adapted from S.M. Sze, Physics of Semiconductor Devices, Wiley, 1981). (**b**) Bipolar transistor subjected to an emitter current pulse and resulting collector current response vs. time (Adapted from S.M. Sze, Physics of Semiconductor Devices, Wiley, 1981). (**c**) Schottky-clamped transistor schematic. d. Data comparing bipolar transistor switch-off times using gold doping and Schottky clamping (Adapted from [1])

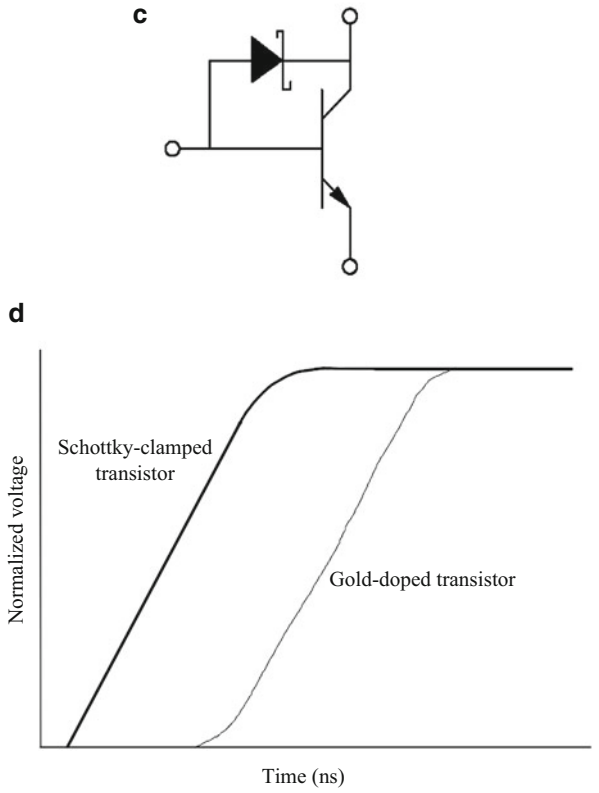


Fig. 3.28 (continued)

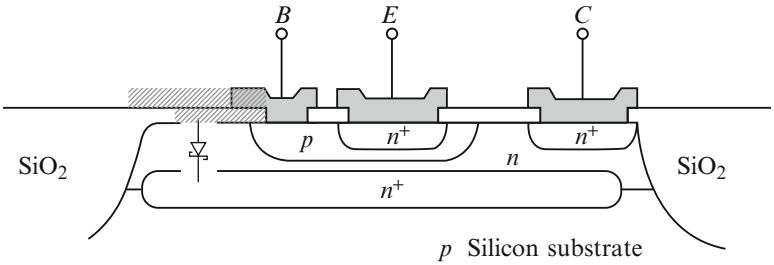


Fig. 3.29 Schottky-clamped transistor superimposed on a conventional *npn* transistor structure. In the Schottky-clamped device the base metal electrode (*gray*) is extended (*hatched region*) in order to create a metal–semiconductor diode with the collector region as indicated

order to aid carrier transport (Fig. 3.31b). GaAs, SiGe, and InP are common semiconductor materials on which HBT devices are based.

Bipolar transistors are very widely used as discrete devices or in analog integrated circuits, where their high drive current and transconductance are advantageous.

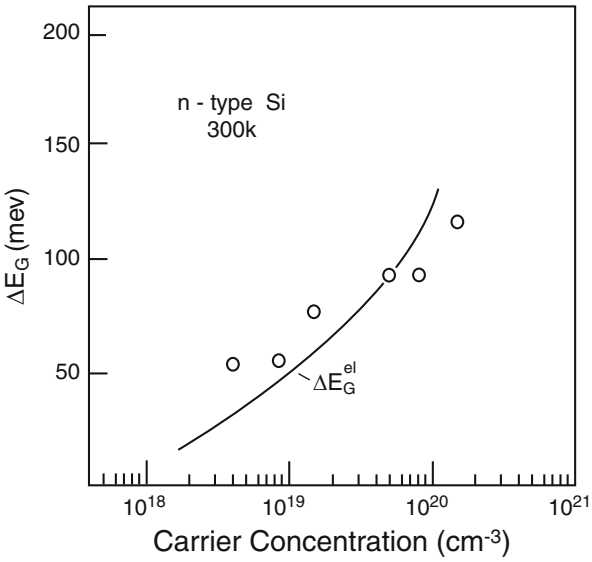


Fig. 3.30 Band gap narrowing in heavily doped silicon. The increased concentration of impurity levels leads near the band edge leads to interactions between energy levels that modify the band structure and effectively reduces the energy band gap. (*Open circles* denote data obtained from optical measurements while the *solid line* is derived from electrical data. *Source*: J. Wagner, J. A. del Alamo, J. Appl. Phys. **63**, 425 (1988))

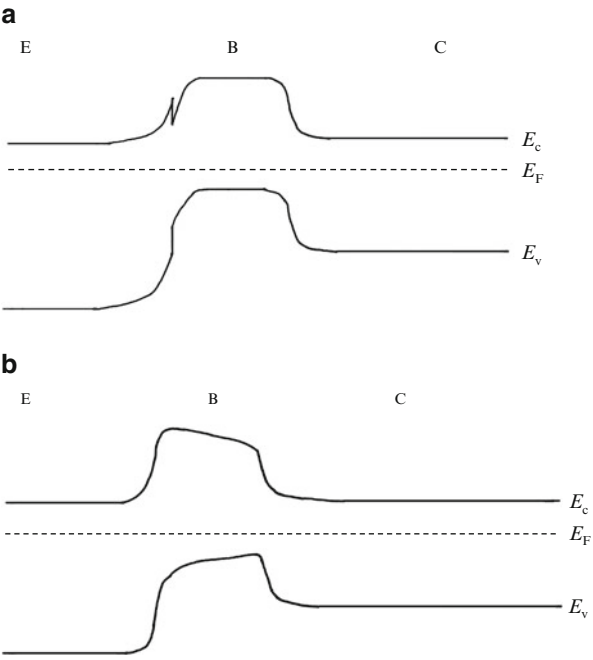


Fig. 3.31 (**a**) Heterojunction bipolar transistor with wide band gap emitter. Under forward-active bias, the larger barrier for holes in the base leads to an increased emitter injection efficiency. (**b**) HBT with a graded base resulting in field-assisted transport of minority carrier electrons across the base. In addition to increasing transistor speed and output current (and therefore gain), this graded-base design also increases the Early voltage since the doping level near the collector is larger, which reduces base-width modulation.

In addition, they can be combined with field effect transistors in hybrid digital integrated circuits known as *BiCMOS* to improve the performance of digital logic or mixed signal applications, which will be discussed further in Chap. 4.

References

1. Muller, R.S., Kamins, T.I.: Device Electronics for Integrated Circuits, 3rd edn. Wiley, New York (2003)
2. Leaver, K.D.: Microelectronic Devices, 2nd edn. Imperial College Press, London (1997)
3. Sze, S.M., Lee, M.K.: Semiconductor Devices Physics and Technology, 3rd edn. Wiley, New York (2012)
4. Ng, K.K.: Complete Guide to Semiconductor Devices, 2nd edn. Wiley Interscience, New York (2002)
5. Schilling, D.L., Belove, C.: Electronic Circuits: Discrete and Integrated. McGraw-Hill, New York (1968)

Problems

1. *Gummel numbers*. Assume the data in Fig. 3.5 was measured on an idealized *npn* transistor structure with a base width $x_B = 0.75$ μm. Find the Gummel number for this transistor. (Hint: use an iterative solution.)
2. *Asymmetric transistor structure*. Find β_F for an *npn* bipolar transistor structure with abrupt step junctions having emitter doping $N_{dE} = 2 \times 10^{18}\text{cm}^{-3}$, base doping $N_{aB} = 10^{17}\text{cm}^{-3}$, and collector doping $N_{dC} = 10^{16}\text{cm}^{-3}$. The emitter width is 5 μm, the base width is 0.5 μm, the collector width is 20 μm, and $A = 10^{-4}\text{cm}^2$. Assume also that $\tau_n = 10^{-7}$ s in the base, $\tau_p = 10^{-7}$ s in the collector, and $\tau_p = 10^{-8}$ s in the emitter. Calculate α_R for this transistor and use it to determine the ratio of the Ebers–Moll model parameters, I_{F0}/I_{R0}.
3. *Ebers–Moll equations*. (1) Use the Ebers–Moll model to calculate the voltage present on the base–emitter junction of an *npn* BJT when the base is open-circuited and a reverse bias exists on the base–collector junction. What is the collector current? Assume $\alpha_R = 0.70$, $I_{R0} = 10^{-13}$ A, and $I_{CB0} = 3.14 \times 10^{-14}$ A. (2) Use the Ebers–Moll model to calculate the voltage present on the base–emitter junction of an *npn* BJT when the emitter is open-circuited and a reverse bias is placed on the base–collector junction. Assume $\alpha_F = 0.985$.
4. *Bipolar transistor design*[21]. Assuming an *npn* bipolar transistor structure with abrupt step junctions and 1 μm emitter width, design the doping levels and base width such that the dc current gain is 100 and the punchthrough voltage is 100 V. $\tau_n = 10^{-6}$ *s* in the base and $\tau_p = 5 \times 10^{-8}$ *s* in the emitter.
5. *Bipolar transistor frequency limit*. If the cutoff frequency of a silicon *npn* bipolar transistor is determined solely by the collector transit time, design the doping levels so that the BJT will operate up to 10 GHz for $V_{BC} = -10$ V.

[21] This problem may be somewhat difficult and/or lengthy.

Chapter 4
Field Effect Transistors

"I looked at what we were doing in integrated circuits at that time (1965), and we made a few circuits and gotten up to 30 components on the most complex chips. . .working on 60. . .and in fact from the days of the original planar transistor, which was 1959, we had about ***doubled every year the amount of components we could put on a chip****."*

G. E. Moore

The idea of using an electric field to modulate the conductivity of a semiconductor was first proposed and patented by Lilienfeld in 1925 (Fig. 4.1). This type of *field effect* phenomenon is used today in various types of field effect transistors (FETs). The first demonstration of a working FET device occurred in 1948 with practical devices appearing around 1953.[1]

We will focus most of our discussion on a type of FET structure based on an oxide–silicon interface, due to its wide-reaching technological importance. Silicon is the second most abundant element in the earth's crust after oxygen and is readily available and inexpensive. It has become the dominant semiconductor material mainly because of the ability to produce, in a compatible manner, both a single crystal and an insulator (SiO_2) that have excellent electrical and mechanical properties. This ability has in turn led to the reliable production of large-scale integrated circuits.

Properties of the oxide–silicon system are fundamental to the performance of integrated circuit devices. Adding a metallic layer, or gate, above the oxide provides an electrode at which the voltage can be fixed, and the resulting three-component, metal–oxide–silicon (MOS) system is important for understanding several classes of devices; in particular the metal–oxide–silicon field effect transistor (MOSFET). The first MOSFET was reported in 1960[2] and has since become a device of major

[1] W. Shockley, G. L. Pearson, Phys. Rev. **74**, 232 (1948); G. C. Dacey, I. M. Ross, Proc. IRE **41**, 970 (1953).

[2] D. Kahng, M. M. Atalla, IRE-AIEE Solid-State Device Res. Conf., Pittsburgh, 1960.

C. Papadopoulos, *Solid-State Electronic Devices: An Introduction*,
Undergraduate Lecture Notes in Physics, DOI 10.1007/978-1-4614-8836-1_4,

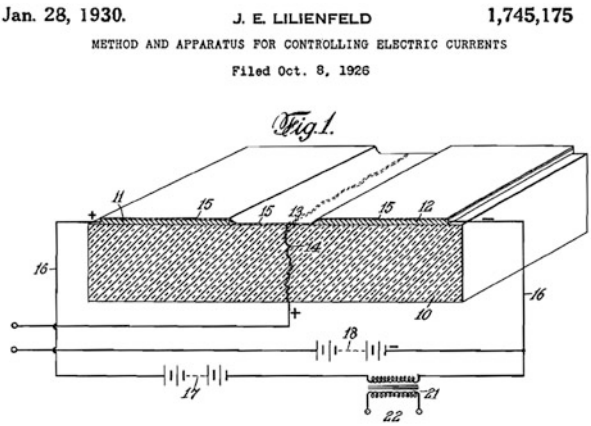


Fig. 4.1 Diagram from early FET patent due to Lilienfeld (excerpt from US patent is shown; also filed in Canada 1 year earlier). The device operation is similar to a modern junction or metal–semiconductor FET (JFET or MESFET) wherein the bias applied to the gate electrode 13 (via connection 14) modulates current flow between 11 and 12 by altering the width of the depletion layer at the semiconductor surface (thin top layer 15)

importance because of the development of dense, very-large-scale integrated (VLSI) circuits that can now accommodate billions of MOSFETs on a single chip, which comprise most of the electronics industry.

4.1 MOS Capacitor System

We first examine the MOS system on its own, which will then help us understand the underlying operation of the MOSFET. Once again we begin by constructing an energy band edge diagram: As we have seen before, the Fermi levels in the different materials are equalized in thermal equilibrium by the transfer of negative charges from materials with higher Fermi levels (smaller work functions) across the interfaces to materials with lower Fermi levels (higher work functions). When the materials in Fig. 4.2a are brought together, electrons are transferred from the aluminum to the silicon to bring the system into equilibrium. For the present, we idealize the MOS system by considering the oxide to be free of charge. Thus the insulator will ideally possess zero mobile charge, and a voltage drop appears across it due to the charge stored on either side.

In equilibrium there is a very thin sheet of positive charge at the surface of the metal and negatively charged acceptors extending into the semiconductor from its surface. Thus the voltage across the MOS system divides across the oxide and the space-charge region near the surface of the silicon. In practice, achieving thermal equilibrium in the MOS system via transfer of charge through an ideal insulating oxide would take a very long time. However, when the MOS structure is fabricated

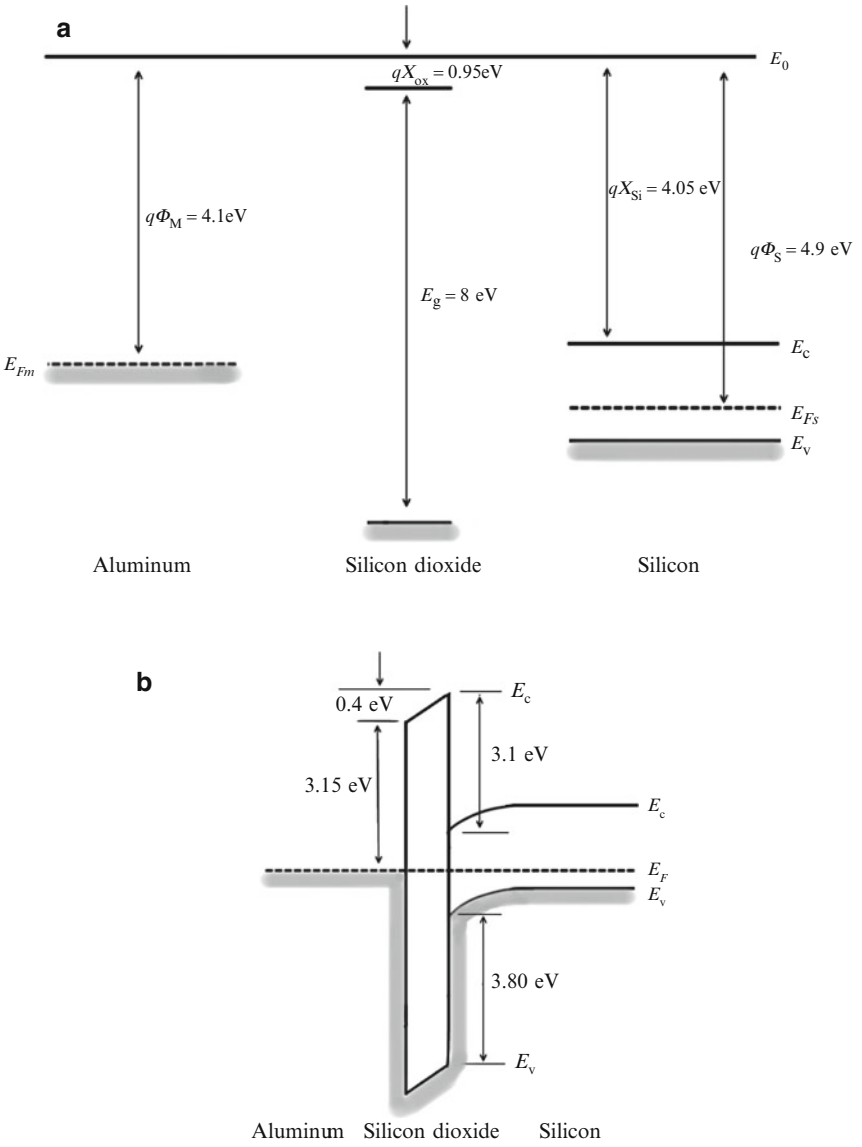


Fig. 4.2 (**a**) MOS energy levels before materials are brought together. In this example, aluminum functions as the gate and the silicon is p-type. (**b**) MOS thermal equilibrium band edge diagram. The voltage dropped across the oxide will depend on its thickness and the total drop across the MOS structure is equal to the difference in work functions between the metal and the semiconductor (similar to a metal–semiconductor junction)

some alternative path for the transfer of charge usually exists, e.g., the aluminum gate electrode may be connected to the silicon substrate or there may be an ohmic conducting path between them. Therefore we can assume that thermal equilibrium exists between the metal and the semiconductor as shown in Fig. 4.2b. In such MOS

band edge diagrams it is convenient to use the oxide "conduction" band edge[3] at the interface as a reference instead of the vacuum level. We can see from this diagram that electrons cannot pass freely in either direction across the oxide because of the potential energy barriers present at the interfaces.

4.1.1 Flat-Band Voltage

In the idealized MOS system the metal and semiconductor can be considered to form two plates of a parallel plate capacitor and this structure is therefore often called a *MOS capacitor*. In thermal equilibrium, the capacitor is charged to a voltage corresponding to the difference between the metal and semiconductor work functions, as we have seen for other systems.

Applying a bias voltage between the metal and silicon changes the amount of charge stored on the capacitor. For the case we considered above, a negative voltage applied to the metal with respect to the silicon opposes the built-in voltage on the capacitor and tends to reduce the charge stored on the plates below its equilibrium value. At some point, the applied voltage will exactly compensate the difference in the work functions of the metal and the semiconductor. The stored charge on the MOS capacitor is then reduced to zero and the fields in the oxide and the semiconductor vanish with the energy bands becoming level or flat. Because of the effect of the applied voltage on the band diagram, the voltage that produces flat energy bands in the oxide and silicon is called the *flat-band voltage* V_{FB} and it equals the difference in the work functions of the metal and silicon:

$$V_{FB} = \Phi_M - \Phi_S \equiv \Phi_{MS} \tag{4.1}$$

The flat-band voltage thus varies with the dopant density in the silicon and the specific metal used as shown for our example in Fig. 4.3.

4.1.2 Accumulation

If we continue to increase the magnitude of the negative voltage applied to the metal past V_{FB} holes are attracted toward the silicon surface and the MOS capacitor begins to store positive charge there (Fig. 4.4a). The positive charge is made up of an increase in the free hole population at the surface. This condition is called (surface) *accumulation* and the resulting region of increased hole population is known as the accumulation layer.

[3] Although the silicon dioxide layer is amorphous, it can be accurately modeled as a large band gap semiconductor.

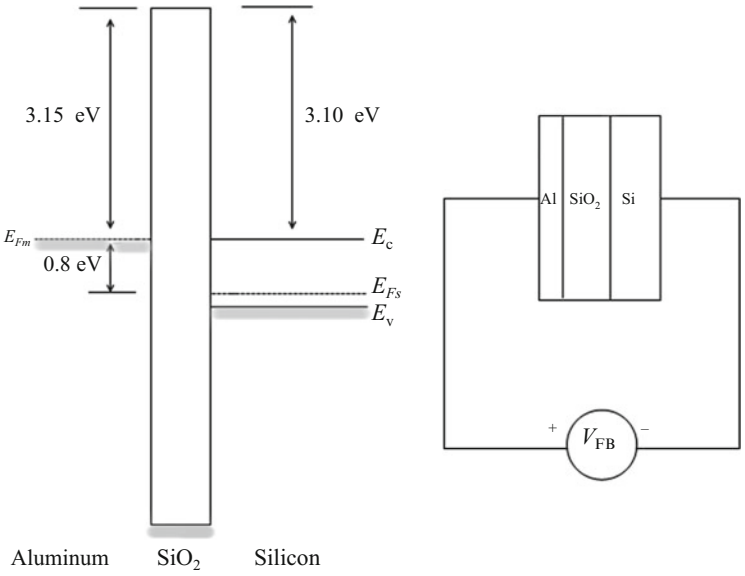


Fig. 4.3 MOS band edge diagram under flat-band condition for the materials shown in Fig. 4.2a. The flat-band voltage, V_{FB}, is equal to −0.8 V for this example

4.1.3 Depletion

Next, we can consider the effect of applying positive bias to the metal with respect to the semiconductor (Fig. 4.4b): A positive gate voltage assists the built-in voltage that we saw was present in thermal equilibrium and the silicon becomes further depleted as holes are repelled from its surface and more acceptor ions are exposed. Because mobile charge carriers are removed from the silicon surface this is known as *surface depletion*.

4.1.4 Inversion

If the positive voltage applied to the metal increases further, however, the increased electric field at the silicon surface will cause the bands to bend by a large amount. At some point this will result in the Fermi level being closer to the conduction band edge than the valence band edge, near the oxide–silicon surface (Fig. 4.4c). At the surface majority carriers have been depleted and generation will exceed recombination. The generated electron–hole pairs are separated by the field—holes are swept away from the surface while electrons move to the oxide–silicon interface where they are confined by the conduction band barrier.

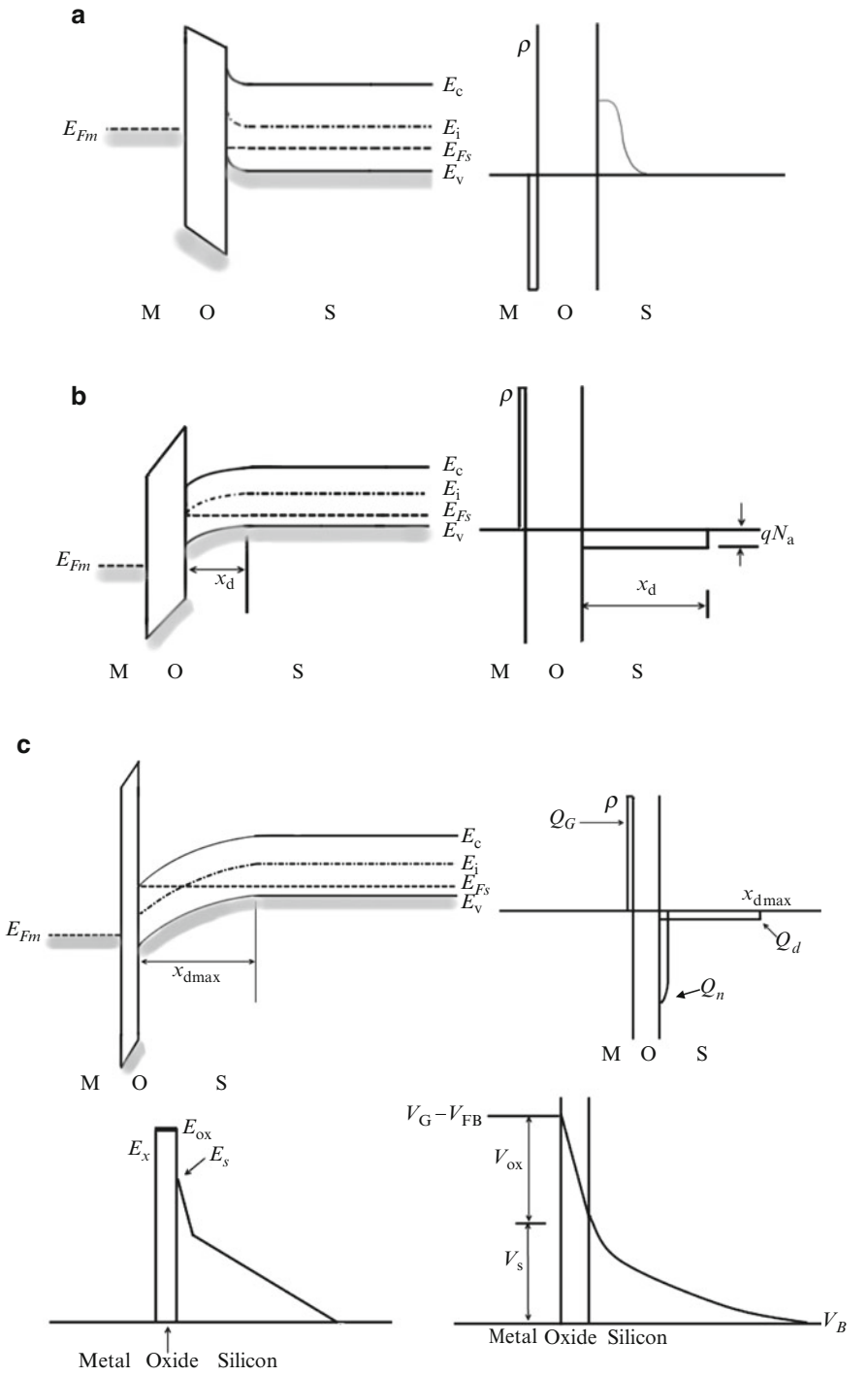


Fig. 4.4 (**a**) MOS capacitor band edge diagram in accumulation condition and resulting charge distribution. (**b**) MOS band edge diagram in depletion and resulting charge distribution. (**c**) MOS band edge diagram in inversion. The charge distribution now consists of both fixed space-charge and mobile electrons and results in the electric field and potential variations shown. (**d**) MOS states and semiconductor charge properties. (Note: A p-type substrate is assumed. For an n-type silicon MOS system the negative charges would instead be positive and vice versa.) Once the MOS system reaches accumulation or inversion the voltage drop occurs mainly across the oxide since both "plates" of the MOS capacitor will have ample charges on either side of the dielectric

d

Condition	Charge in semiconductor
Flat band	No charge–neutral
Accumulation	Free holes near interface
Depletion	Fixed space-charge (depletion layer–negative)
Inversion	Fixed space-charge (negative) plus free electrons near interface

Fig. 4.4 (continued)

In terms of carrier densities, this means that there are now more electrons than holes at the surface of the (p-type) silicon, which is referred to as an *inversion layer*. The applied voltage has essentially induced a *pn* junction near the surface of the semiconductor.

The degree of inversion is classified according to that the amount E_i is below E_F at the surface. When this amount is small, the electron density in the inversion layer is low ($\sim n_i$) and the MOS system is said to be biased in the *weak inversion* regime. On the other hand, when the Fermi level is closer to the conduction band at the surface than it is to the valence band in the bulk, the system is in the *strong inversion* region. The dividing point between the two inversion regimes is usually taken when the electron density at the surface equals the acceptor concentration in the bulk.

Once the silicon is biased into strong inversion, the free electron density at the surface is nearly an exponential function of the potential at the surface and the surface potential will change very little with increasing gate voltage after the inversion layer forms. The charge induced by the additional gate voltage is virtually all free electron charge just below the oxide–silicon interface and thus the total potential drop across the depletion region and the depletion layer width in the silicon are both relatively constant after the surface is inverted.

Qualitatively, the electric field and potential variations in the MOS structure biased into inversion will have the form shown in Fig. 4.4c. The different states of the MOS system and their properties are summarized in Fig. 4.4d.

4.1.5 Model for Charges in the Silicon Substrate

Thus far the qualitative discussion of surface charge in the MOS system emphasized the important effect the gate voltage has in determining the properties of the silicon surface. Although the system is basically a capacitor, the various forms that the surface charge can take produce very significant differences in the electrical

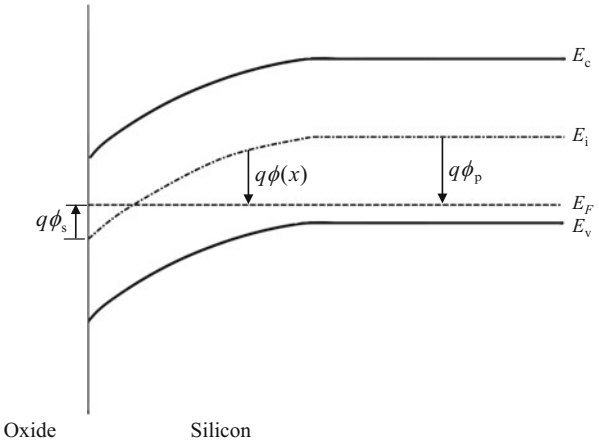


Fig. 4.5 Definition of potential in the silicon substrate of the MOS system

properties of the silicon surface. For example, the surface can be highly conductive and electrically connected to free carriers in the bulk when it is biased in accumulation; highly resistive when it is biased in depletion; or highly conductive but disconnected from the bulk when it is biased in inversion.

To obtain a more quantitative model of the charges induced near the silicon surface, we first define the potential with respect to the Fermi level in the silicon as

$$\phi(x) = \frac{1}{q}[E_{\mathrm{F}} - E_{\mathrm{i}}(x)] \tag{4.2}$$

As shown in Fig. 4.5 the potential in the neutral bulk is negative because the material is p-type and E_{F} is less than E_{i}. At the surface, the potential is defined as

$$\phi(0) = \phi_s = \frac{1}{q}[E_{\mathrm{F}} - E_{\mathrm{i}}(0)] \tag{4.3}$$

The carrier densities are related to $\phi(x)$ by the usual expressions[4]:

$$p = n_{\mathrm{i}} \exp\left(-\frac{q\phi}{k_{\mathrm{B}}T}\right), n = n_{\mathrm{i}} \exp\left(\frac{q\phi}{k_{\mathrm{B}}T}\right) \tag{4.4}$$

Using these equations we can express the free carrier densities at the surface, n_{s} and p_{s}, in terms of the total potential drop, $(\phi_{\mathrm{s}} - \phi_{\mathrm{p}})$, across the depletion region as

[4] See Appendix A, Eq. (A.29).

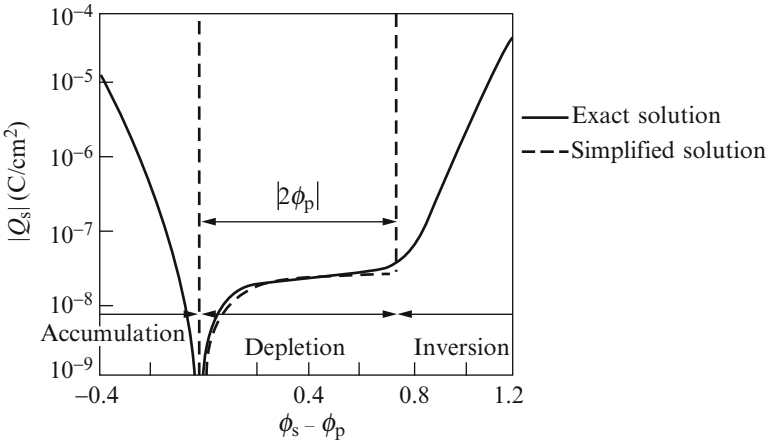


Fig. 4.6 Charge density vs. potential drop in silicon for a MOS structure (Adapted from [2]). In the simplified model, three regimes of operation are identified with sharp boundaries

$$p_s = N_a \exp\left[\frac{q(\phi_p - \phi_s)}{k_B T}\right]$$
$$n_s = \frac{n_i^2}{N_a} \exp\left[\frac{q(\phi_s - \phi_p)}{k_B T}\right] \tag{4.5}$$

These equations show that the densities of free surface charges that pile up against the oxide–silicon interface in accumulation or inversion will depend exponentially on the amount of local band bending, as required by thermal statistics. The exact solution[5] for the total silicon space-charge density Q_s, including both the free carriers and depletion region charge, is shown in Fig. 4.6 along with an approximate model that will allow more straightforward analytical results to be obtained: For the simplified model (dotted lines, Fig. 4.6), we assume that the onset of accumulation occurs at the flat-band condition and that inversion starts when the potential drop reaches a value of $2\phi_p$. In addition, we assume the only charges present in depletion are the fixed ionized impurities (i.e., the usual depletion approximation). These are usually very good assumptions since the free carrier concentrations increase very rapidly (exponentially) in inversion or accumulation. Finally, the inversion and accumulation layers are assumed to have zero thickness so that the free charges form a thin sheet of charge at the silicon surface, and there is no additional band bending within this layer. These assumptions, sometimes referred to as the *charge sheet model*, yield results with adequate accuracy for most purposes.

[5] C. G. B. Garrett, W. H. Brattain, Phys. Rev. **99**, 376 (1955).

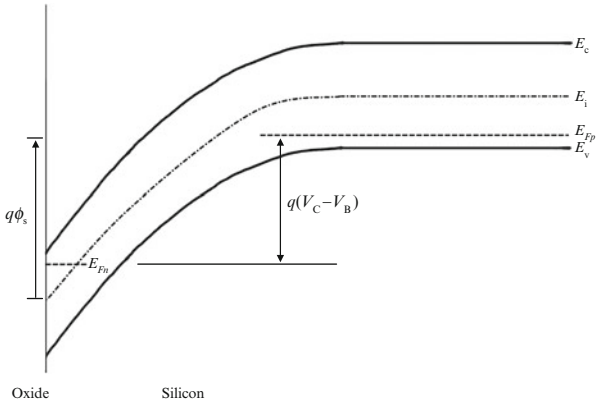


Fig. 4.7 Band edge diagram depicting the effect of biasing the n-type inversion layer (or channel) of a MOS capacitor positively with respect to the p-type bulk

With this idealized model, we can relate the surface potential to the depletion layer width by using the depletion approximation, exactly as we did in Chap. 2 for the Schottky diode (and one-sided *pn* diode):

$$x_d = \sqrt{\frac{2\varepsilon_s|\phi_s - \phi_p|}{qN_a}} \tag{4.6}$$

and the space-charge density (per unit area) in depletion is therefore

$$Q_d = -qN_a x_d \tag{4.7}$$

The maximum depletion width that can be attained within this model occurs at the onset of inversion,

$$x_{dmax} = \sqrt{\frac{4\varepsilon_s|\phi_p|}{qN_a}} \tag{4.8a}$$

and therefore,

$$Q_{dmax} = -qN_a x_{dmax} = -\sqrt{4\varepsilon_s qN_a|\phi_p|} \tag{4.8b}$$

4.1.5.1 Effect of substrate bias

We have seen that once the MOS system is biased into inversion a *pn* junction exists between the surface and the bulk of the silicon (the substrate). If a reverse bias is applied between the induced surface n-region and the bulk, the depletion width increases as illustrated in Fig. 4.7. The reverse bias (V_C–V_B) is applied between the inversion layer (or *channel*) and the substrate (or *bulk*). The net result of applying a

bias between the channel and the bulk is to make the maximum width of the depletion region larger compared to Eq. (4.8a):

$$x_{\mathrm{dmax}} = \sqrt{\frac{2\varepsilon_{\mathrm{s}}\left(2\left|\phi_{\mathrm{p}}\right| + V_{\mathrm{C}} - V_{\mathrm{B}}\right)}{qN_{\mathrm{a}}}} \tag{4.9a}$$

and also

$$Q_{\mathrm{dmax}} = -\sqrt{2\varepsilon_{\mathrm{s}} qN_{\mathrm{a}}\left(2\left|\phi_{\mathrm{p}}\right| + V_{\mathrm{C}} - V_{\mathrm{B}}\right)} \tag{4.9b}$$

4.1.5.2 Threshold Voltage

We have seen that flat band (V_{G}–$V_{\mathrm{B}} = V_{\mathrm{FB}}$) corresponds to charge neutrality in the silicon. Therefore, we can say that

$$\left(V_{\mathrm{G}} - V_{\mathrm{B}}\right) - V_{\mathrm{FB}}$$

is the effective voltage that charges the MOS capacitor. To express the capacitor charge in terms of applied voltage, we carry out the following analysis:

The voltage across the MOS capacitor is the sum of a voltage drop across the oxide and a drop in the silicon:

$$V_{\mathrm{G}} - V_{\mathrm{B}} - V_{\mathrm{FB}} = V_{\mathrm{ox}} + \phi_{\mathrm{s}} - \phi_{\mathrm{p}} \tag{4.10}$$

The field in the insulating oxide is constant in the absence of oxide charge and we can therefore write

$$E_{\mathrm{ox}} = V_{\mathrm{ox}}/x_{\mathrm{ox}} = \left[\left(V_G - V_{\mathrm{B}} - V_{\mathrm{FB}}\right) - \left(\phi_{\mathrm{s}} - \phi_{\mathrm{p}}\right)\right]/x_{\mathrm{ox}} \tag{4.11}$$

where x_{ox} is the oxide thickness. The electric field on either side of the oxide/silicon boundary must satisfy[6]

$$\varepsilon_{\mathrm{ox}} E_{\mathrm{ox}} = \varepsilon_{\mathrm{s}} E_{\mathrm{s0}} \tag{4.12}$$

Combining Eq. (4.12) with Eq. (4.11), and using as the definition of oxide capacitance (per unit area) $C_{\mathrm{ox}} = \varepsilon_{\mathrm{ox}}/x_{\mathrm{ox}}$, gives

$$\varepsilon_{\mathrm{s}} E_{\mathrm{s0}} = C_{\mathrm{ox}}\left[\left(V_{\mathrm{G}} - V_{\mathrm{B}} - V_{\mathrm{FB}}\right) - \left(\phi_{\mathrm{s}} - \phi_{\mathrm{p}}\right)\right] \tag{4.13}$$

[6] Recall, this condition arises from the conditions set by Maxwell's equations on the displacement field ($D_x = \varepsilon_0 \varepsilon_{\mathrm{r}} E_x$) normal to an interface, which in the absence of oxide surface charge must be continuous.

Now, by applying the integral form of Gauss' law to a volume extending from just inside the silicon at the oxide/silicon interface to the field-free bulk region we can write

$$-\varepsilon_s E_{s0} = Q_s = Q_n + Q_d \tag{4.14}$$

where Q_s is the total charge induced in the semiconductor under inversion, which is composed of the mobile electron charge Q_n and the depletion region charge Q_d (all per unit area). This allows us to write an expression for the mobile charge in the MOS system as a function of applied gate voltage:

$$Q_n = -C_{ox}\left[(V_G - V_B - V_{FB}) - (\phi_s - \phi_p)\right] - Q_d \tag{4.15}$$

In (strong) inversion this becomes[7]

$$Q_n = -C_{ox}\left(V_G - V_{FB} - V_C - 2|\phi_p|\right) + \sqrt{2\varepsilon_s q N_a\left(2|\phi_p| + V_C - V_B\right)} \tag{4.16a}$$

When no bias is applied between the channel and the bulk we get the simpler expression:

$$Q_n = -C_{ox}\left(V_G - V_{FB} - V_B - 2|\phi_p|\right) + \sqrt{4\varepsilon_s q N_a |\phi_p|} \tag{4.16b}$$

If we set either one of Eqs. (4.16a) and (4.16b) equal to zero we can directly express the gate voltage necessary to induce a conducting channel at the surface of the semiconductor under inversion. This voltage is known as the *threshold voltage* V_T :

$$V_T = V_{FB} + V_C + 2|\phi_p| + \frac{1}{C_{ox}}\sqrt{2\varepsilon_s q N_a\left(2|\phi_p| + V_C - V_B\right)} \tag{4.17}$$

Once the MOS system is in strong inversion, the mobile charge formed at the surface can also be expressed simply in terms of the difference between the applied gate voltage and the threshold voltage by analogy with the parallel plate capacitor:

$$Q_n = -C_{ox}(V_G - V_T) \tag{4.18}$$

This result is based on the initial simplifying assumption that no mobile charge exists at the surface when V_G is equal to V_T, which is not exactly correct (see Fig. 4.6). Thus Eq. (4.18) is less accurate when V_G is close to V_T.

[7] Equations (4.16a) and (4.16b) are obtained by substituting for the potential drop and space-charge density in the silicon at the onset of inversion (see Eqs. (4.8a), (4.8b), (4.9a), and (4.9b) and Fig. 4.6).

The analogous treatment for a MOS system consisting of an n-type substrate leads to

$$V_T = V_{FB} + V_C - 2|\phi_n| - \frac{1}{C_{ox}}\sqrt{2\varepsilon_s q N_d\left(2|\phi_n| + V_B - V_C\right)} \tag{4.19}$$

and[8]

$$Q_p = C_{ox}(V_T - V_G) \tag{4.20}$$

for the charge (now positive mobile holes) in strong inversion.

4.1.6 Deviations from Ideal Behavior

4.1.6.1 Oxide and Interface Charge

The MOS theory presented to this point has not considered the influence of charge within the oxide and at the oxide–silicon interface. Although the silicon/silicon dioxide system has been refined to contain a very low charge density compared to other material systems the presence of these charges still needs to be considered. Ionic impurities/defects within the oxide layer and defects at the oxide–silicon interface introduce additional charge densities to the MOS system: Fixed charge at the interface is caused by a thin layer of nonstoichiometric silicon dioxide (SiO_x) and is positive. The oxide charge density is distributed in traps throughout the oxide layer and can be either positive or negative.

The different sources of oxide charge will induce opposite charges at the semiconductor surface and thus alter the flat-band voltage of the ideal MOS system to

$$V_{FB} = \Phi_{MS} - \frac{Q_f}{C_{ox}} - \frac{1}{C_{ox}}\int_0^{x_{ox}} \frac{x}{x_{ox}}\rho(x)dx \tag{4.21}$$

where Q_f is the fixed interface charge density and $\rho(x)$ is the distributed oxide charge. The reduction of defects and trap densities in modern MOS devices is one of the main reasons they became commercially viable.

Example 4.1: MOS Threshold Voltage Calculation Find the threshold voltage in a 0.5 Ω-cm p-type silicon MOS system characterized by (1) $p+$ polysilicon gate, (2) 10 nm silicon dioxide, (3) the oxide is free of charge except for a surface charge density $(Q_f/q) = 10^{11}\text{cm}^{-2}$, and (4) $V_C = V_B = 0$.

[8] Note that in this case the gate voltage will be negative in order to induce a positive layer of mobile charge in the n-type substrate.

Looking up the required data from Appendix B,

0.5 Ω-cm p-type Si has $N_a = 3 \times 10^{16}\text{cm}^{-3}$, and

$$|\phi_p| = \frac{k_B T}{q} \ln\left(\frac{N_a}{n_i}\right) = 0.377\,\text{V}$$

$$V_{FB} = \Phi_{MS} - \frac{Q_f}{C_{ox}} = [(4.05 + 1.12) - (4.05 + 0.56 + 0.377)] - \frac{1.6 \times 10^{-19} \cdot 10^{11} \cdot 10 \times 10^{-7}}{3.9 \cdot 8.85 \times 10^{-14}} = 0.1366\,\text{V}$$

$$\frac{|Q_{dmax}|}{C_{ox}} = \frac{\sqrt{4\varepsilon_s q N_a |\phi_p|}}{C_{ox}} = \frac{10 \times 10^{-7}}{3.9 \cdot 8.85 \times 10^{-14}} \sqrt{4 \cdot 11.7 \cdot 8.85 \times 10^{-14} \cdot 1.6 \times 10^{-19} \cdot 3 \times 10^{16} \cdot 0.377} = 0.251\,\text{V}$$

The threshold voltage is thus given by

$$V_T = V_{FB} + 2|\phi_p| + \frac{|Q_{dmax}|}{C_{ox}} = 0.1366 + 2 \times 0.377 + 0.251 = 1.14\,\text{V}$$

4.1.7 Capacitance of the MOS structure

The small-signal capacitance of the MOS system,

$$C = \frac{dQ}{dV}$$

can depend strongly on the frequency of the applied voltage.

First, assume the MOS system is subjected to a dc bias V_G that causes the silicon surface to be accumulated. For a p-type silicon sample this corresponds to a negative applied voltage as we saw earlier, where the gate voltage pulls the excess holes close to the oxide interface at the silicon surface. If a small ac voltage v_G is superimposed on the dc bias, it will cause small variations in the charges stored on the gate and at the silicon surface. The capacitance in this case will be nearly equal to that of the oxide itself since the spatial extent of the accumulation layer is assumed small, i.e.,

$$C_{ox} = \frac{\varepsilon_{ox}}{x_{ox}} \tag{4.22}$$

When the gate voltage becomes more positive than the flat-band voltage the system moves into depletion. The overall capacitance now becomes a series connection of oxide capacitance and surface depletion capacitance:

$$C = \frac{1}{1/C_{ox} + 1/C_s} = \frac{1}{x_{ox}/\varepsilon_{ox} + x_d/\varepsilon_s} \tag{4.23}$$[9]

Once the gate bias is increased enough to invert the surface, mobile charge can be located very close to the oxide–silicon interface. However this charge is now due to minority carriers that must be thermally generated. Thus, the population of the inversion layer can change only as fast as carriers can be generated within the depletion region near the surface. If both the dc gate bias voltage and the small-signal voltage are changed very slowly so that the silicon can always approach equilibrium, the charges on the gate and those in the silicon are separated only by the gate oxide and the MOS capacitance approaches C_{ox}.

The characteristic time to form an inversion layer at the surface of a modern MOS system is typically on the order of seconds and thus the small-signal voltage must be changed very slowly to observe the low-frequency (LF) C–V curve. On the other hand, if the dc bias is varied slowly while the ac signal is changed rapidly, the inversion layer can follow the dc bias but not the ac bias. The change in charge then corresponds to the movement of holes in the depletion region and the capacitance is given by

$$C_{min} = \frac{1}{x_{ox}/\varepsilon_{ox} + x_{dmax}/\varepsilon_s} \tag{4.24}$$

This high-frequency (HF) capacitance remains constant as the dc gate voltage is increased further.

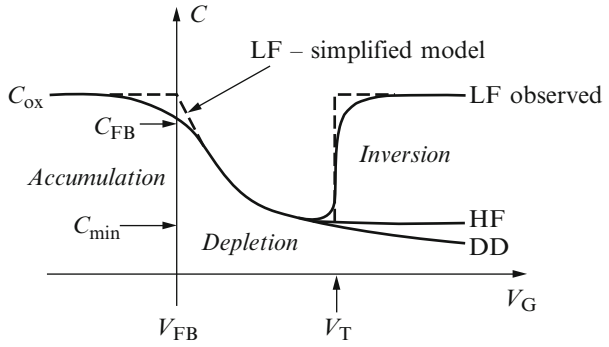
Finally, deep-depletion (DD) capacitance corresponds to the situation in which both the dc gate bias and small-signal voltage vary at a rate faster than the inversion layer can respond. Because the inversion layer cannot form in this case, the depletion region instead becomes wider than x_{dmax} (Fig. 4.8b) and the capacitance drops below C_{min} as shown in Fig. 4.8a.

Note the above analysis implicitly assumed that the MOS system was free of oxide charges. In practice, these charges modify the observed C–V behavior, which is used for MOS technology process monitoring in order to extract information about the nature and amount of oxide charge.

Panel 4.1: Charge-Coupled Devices The MOS capacitor is very useful for several applications including high-resolution imaging arrays. Exposing a MOS system to light when it is biased into deep depletion will generate electron–hole pairs in the silicon: The electric field at the surface will then separate the electrons and holes, resulting in *photogenerated* charge in the capacitor (Fig. 4.9).

[9] Recall that capacitors in series sum like resistors in parallel.

a



b

Bias Condition	Charge Equilibrium	Charge Distribution	Equivalent Circuit	Comment
Low frequency (LF)	dc equilibrium ac equilibrium	$x_d = x_{dmax}$, x, SiO_2, δQ_n	C_{ox}	$\delta Q_s = \delta Q_n$ ac coupling only to surface charge
High frequency (HF)	dc equilibrium ac nonequilibrium	$x_d = x_{dmax}$, SiO_2, δQ_d	C_{ox} ε_s / x_{dmax}	$\delta Q_s = \delta Q_d$ ac coupling only to bulk charge
Deep depletion (DD)	dc nonequilibrium ac nonequilibrium	$x_d > x_{dmax}$, SiO_2, δQ_d	C_{ox} ε_s / x_d	both ac and dc coupling to Q_d $x_d > x_{dmax}$

Fig. 4.8 (**a**) Ideal (no oxide charge) MOS *C*–*V* curves for different rates of varying the applied voltage. (**b**) Table defining different biasing conditions and resulting charge distributions along with corresponding capacitance

An array of such MOS capacitor pixels can be used to capture images as illustrated in Fig. 4.9: Charge-coupled device[10] (CCD) arrays with millions of pixels are widely used for imaging, such as in digital cameras, television cameras, etc. For proper operation of a CCD array the photogenerated charge must be stored and transferred out of the array before the background thermal charge generation becomes significant. This can be achieved by using a high-quality oxide–silicon interface that has a low density of generation–recombination sites (i.e., a long minority carrier lifetime).

[10] W. S. Boyle, G. E. Smith, Bell Syst. Tech. J. **49**, 587 (1970).

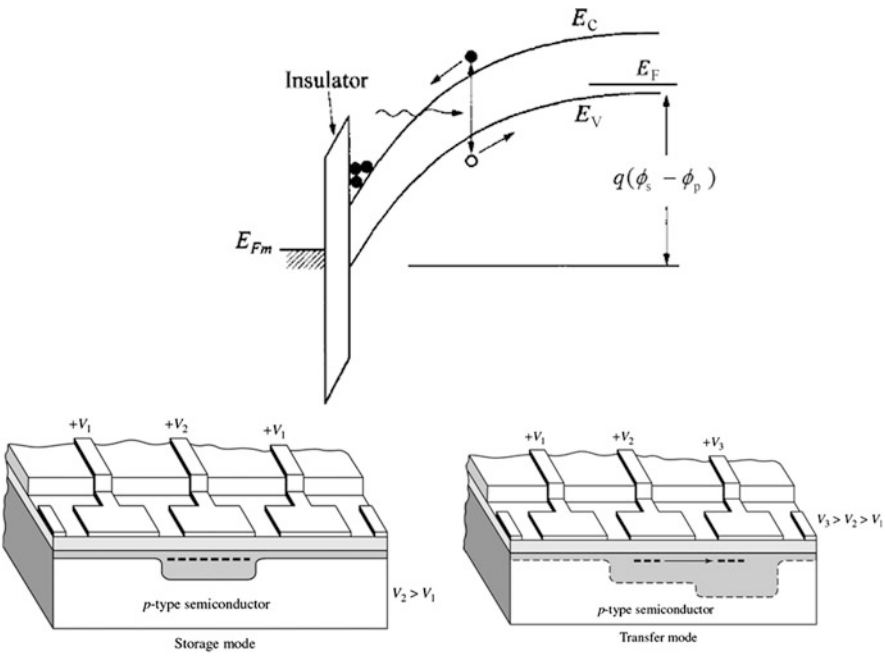


Fig. 4.9 Charge-coupled devices. The MOS system is driven into deep depletion in order to maximize photogenerated charge collection near the interface upon light exposure. Once collected, the charge is read out using an array of MOS capacitors that are tightly coupled as illustrated. For imaging applications, the CCD gates must be thin enough to let the light through (After [1] and [2])

4.2 MOSFETs

To go from the MOS structure to a transistor we add two contact regions on either side of the oxide–semiconductor interface. These source and drain regions become electrically connected to each other when an inversion layer or conducting channel is formed in the MOS capacitor via the gate voltage, which is the basic mechanism of transistor operation in the MOSFET.

The two types of MOS transistors are *n*-channel MOSFETs (in which the conducting carriers are electrons) and p-channel MOSFETs (where the conducting carriers are holes). These are sometimes called NMOSFETs and PMOSFETs (or simply NMOS and PMOS). *n*-channel MOSFETs are built in p-type silicon substrates so that the conducting channels are isolated from the substrate and nearby devices (since an effective *pn* junction is formed), and p-channel MOSFETs are built in n-type silicon for the same reason. For *n*-channel MOSFETs, positive gate voltages that are sufficiently large create a conducting channel, whereas for *p*-channel MOSFETs, negative gate voltages of sufficient magnitude produce a conducting channel.

We will focus our analysis on the *n*-channel MOSFET. However, the same theory will also apply to *p*-channel devices with suitable changes in the signs of appropriate parameters.

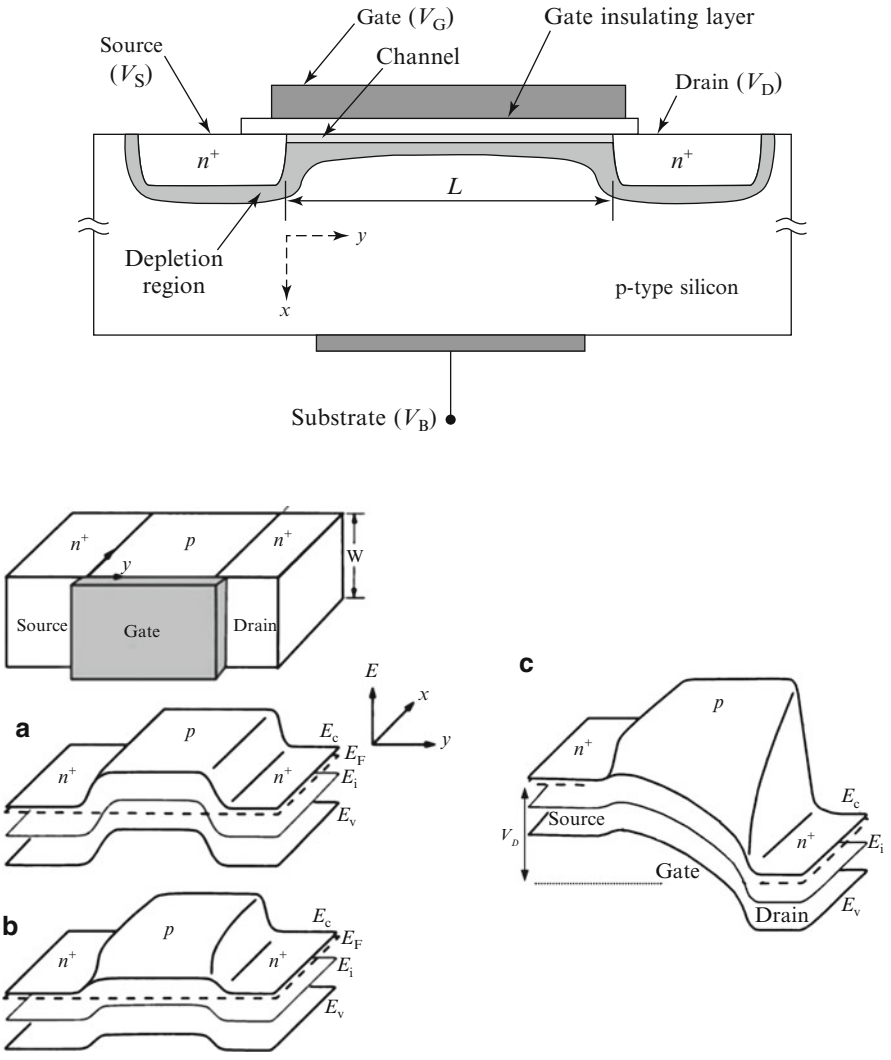


Fig. 4.10 *n*-channel MOSFET device structure schematic. For proper operation the source and drain voltages must always be greater than or equal to that of the p-type substrate (in order for the *pn* junctions to always be reverse biased). When V_G is above threshold the resulting inversion layer on the MOS capacitor creates a conducting channel that connects the source and drain. Two-dimensional band edge diagrams illustrate the MOSFET channel in thermal equilibrium (**a**), above threshold (**b**), and above threshold for large finite drain bias (**c**) (cf. Fig. 4.5) (Adapted from [1])

The basic *n*-channel MOSFET structure is shown in Fig. 4.10. When the surface is inverted and a voltage is applied between the source and drain carriers can enter the channel at the source and leave at the drain. This results in current flow from drain to source in *n*-channel MOSFETs (and from source to drain in *p*-channel MOSFETs). MOSFETs can also be fabricated in which the channel is inverted

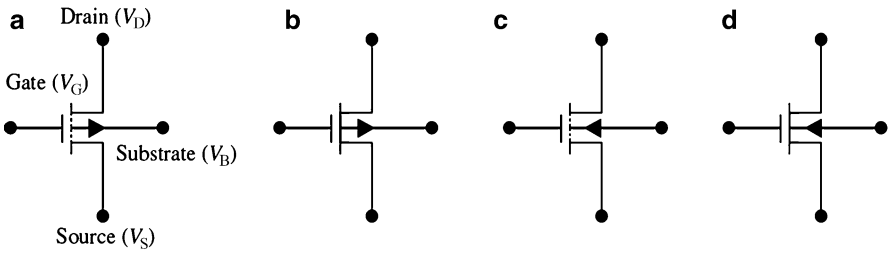


Fig. 4.11 Electrical symbols for MOSFETs: (**a**) *p*-channel enhancement, (**b**) *p*-channel depletion, (**c**) *n*-channel enhancement, (**d**) *n*-channel depletion. (Often a simplified symbol is used with the specific device type implied from the context)

when the gate-to-source voltage is zero. The drain current in this type of MOSFET can be *reduced* by changing the gate-to-source voltage and hence it is called a *depletion-mode* MOSFET. MOSFETs in which the channel region is not inverted at $V_{GS} = 0$, which we have been implicitly considering thus far, are called *enhancement-mode* MOSFETs; enhancement-mode MOSFETs are more frequently used than depletion-mode devices due to their lower standby power dissipation.

Voltages at four terminals affect the behavior of a MOSFET (Fig. 4.11). To develop the basic theory of the MOSFET we initially consider the case with the source and bulk/substrate (also known as the body or back) at the same voltage ($V_S = V_B$). When the voltage applied to the gate of an *n*-channel MOSFET is lower than the threshold voltage V_T, the surface region of the substrate between the source and drain is either accumulated (many holes are present) or depleted of mobile carriers. In both cases, (to first order) virtually no conduction is possible between the n-type source and drain regions and the MOSFET is OFF. Increasing the gate voltage beyond V_T strongly increases the density of mobile channel electrons and thus reduces the source–drain resistance bringing the MOSFET to its ON state. The threshold voltage is thus taken as the transition voltage at which the MOSFET changes from OFF to ON when it is used as a switch.

4.2.1 Long-Channel Theory

When the transistor is ON (i.e., in strong inversion) the channel electrons move mainly by drift,[11] and the drain current can be written as

$$I_D = WQ_n(y)v(y) \tag{4.25}$$[12]

[11] When in the ON state the channel is essentially a resistor.

[12] This is a one-dimensional version of the standard equation for current density $J = qnv$.

where $Q_n(y)$ is the inversion charge per unit area at a position y in the channel and $v(y)$ is the velocity of carriers at that position (see Fig. 4.10). W is the channel width (in the z-direction). At low drain voltage V_D, the drift velocity is given by

$$v(y) = -\mu_n E_y \tag{4.26a}$$

where μ_n is the mobility of the carriers *in the channel*[13] and

$$E_y = -\frac{\partial V(y)}{\partial y} \tag{4.26b}$$

Using Eqs. (4.26a and 4.26b) the drain current can therefore be rewritten as

$$I_D = WQ_n(y)\mu_n \partial V(y)/\partial y \tag{4.27}$$

If we assume that both the mobile channel charge and the bulk depletion charge are controlled by the vertical field only, then the theory we developed for the MOS capacitor in Sect. 4.1 can be applied in the x- (vertical) direction to calculate the inversion charge density. In other words, the field in the y-direction is taken to be much less than that in the x-direction, known as the *gradual channel approximation*:

$$Q_n(y) = -C_{ox}[V_G - V_T - V(y)] \tag{4.28}$$

Substituting this expression into Eq. (4.27) and integrating along the channel from the source ($y = 0$), which is taken as the voltage reference ($V_S = 0$), to the drain ($y = L$) where $V(y) = V_D$, gives

$$\int_0^L I_D dy = \mu_n W C_{ox} \int_0^{V_D} [V_G - V_T - V(y)]dV \tag{4.29a}$$

which can be solved to obtain an equation for the dc drain current:

$$I_D = \mu_n C_{ox} \frac{W}{L}\left[\left(V_G - V_T - \frac{1}{2}V_D\right)V_D\right] \tag{4.29b}$$

known as the *long-channel MOSFET equation*. If we consider this equation with $V_G > V_T$ and increase V_D from zero (Fig. 4.12a) while V_G stays fixed, we see that drain current initially increases linearly with increasing drain voltage, but that the slope decreases when $V_D/2$ becomes appreciable compared to $(V_G - V_T)$. At sufficiently high V_D, the slope becomes negative according to this equation—this is not physical because Eq. (4.28),

[13] The channel carrier mobility is reduced from bulk carrier mobility due to increased scattering at the oxide interface (see Appendix B).

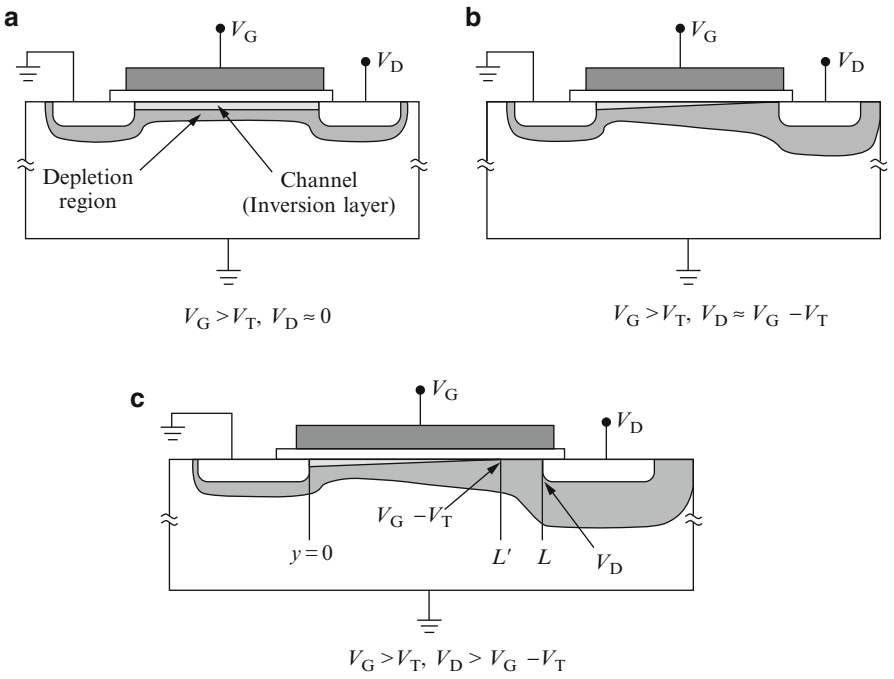


Fig. 4.12 MOSFET channel at low drain voltage (**a**), pinch-off (**b**), and beyond pinch-off point (**c**)

$$Q_n(y) = -C_{ox}[V_G - V_T - V(y)]$$

which we used to derive the long-channel equation is only valid when Q_n is negative, i.e.,

$$V(y) < (V_G - V_T) \tag{4.30a}$$

$V(y)$ has its maximum value V_D at the drain; hence the long-channel MOSFET equation is only valid when

$$V_D \leq (V_G - V_T) \tag{4.30b}$$

When $V_D = (V_G - V_T)$ the mobile electron density in the channel at $y = L$ decreases to zero and the channel becomes *pinched off* as illustrated in Fig. 4.12b. When the drain voltage exceeds the pinch-off voltage, the conducting channel becomes separated from the drain and a voltage drop equal to $[V_D - (V_G - V_T)]$ exists between the doped region of the drain and the location in the channel where the inversion charge becomes zero (the pinch-off point) (Fig. 4.12c).

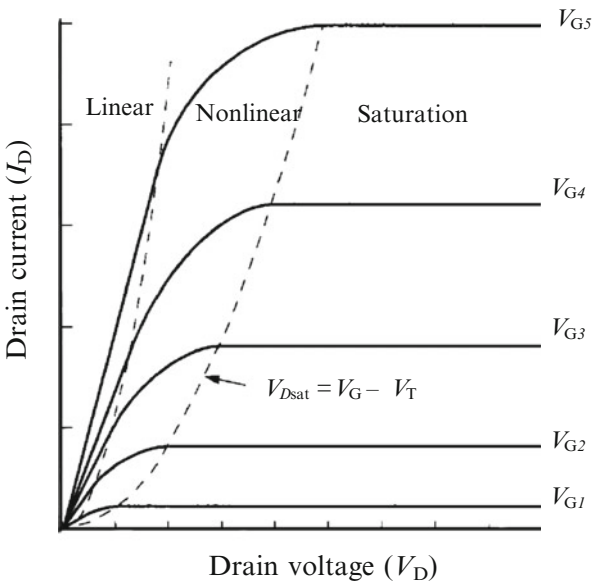


Fig. 4.13 MOSFET *I–V* characteristics based on ideal long-channel theory for different values of gate voltage (Gate voltage increases monotonically from bottom to top)

The MOSFET is said to be operating in the *saturation region* when $V_D > V_{Dsat}$, where $V_{Dsat} = V_G - V_T$. Thus in saturation, according to our model, the drain current no longer increases with increasing drain voltage. The current in saturation is given by

$$I_{Dsat} = I_D(V_{Dsat}) = \mu_n C_{ox} \frac{W}{2L}(V_G - V_T)^2 \tag{4.31}$$

sometimes referred to as "square law" theory. Combining this equation with the long-channel MOSFET equation (Eq. 4.29b) we obtain expressions that describe the set of I_D versus V_D curves for different gate voltages shown in Fig. 4.13: The current increases according to the MOSFET equation until it reaches I_{Dsat} at $V_D = V_{Dsat}$ and then remains constant (to first order).

Since current must be constant along the channel while the density of mobile carriers Q_n carrying that current decreases, the charge velocity $v(y)$ must increase continuously along the channel [see Eq. (4.25)]. This means that in our model the velocity needs to become infinite at the pinch-off point where Q_n becomes zero. This physical impossibility does not invalidate the long-channel model because the drain current is determined by the delivery rate of electrons along the inverted channel between the source and the pinch-off point—once the carriers leave the channel they "fall" through the energy drop in the high-field space-charge region near the drain. The MOSFET water analogy illustrated in Fig. 4.14 is useful for visualizing this behavior.

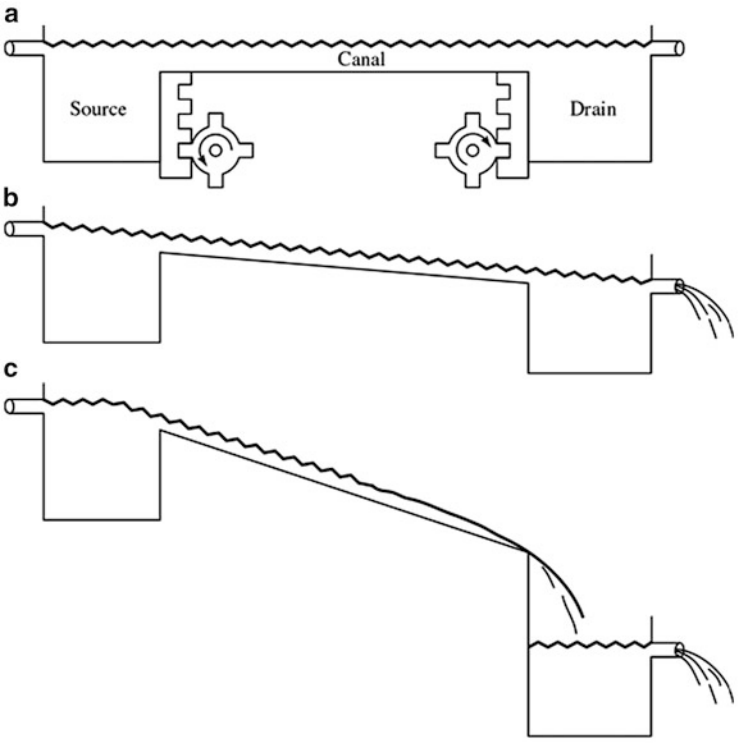


Fig. 4.14 MOSFET water analogy. The canal can be thought of as the channel with the source and drain being water reservoirs. By adjusting the two gears a gravitational potential difference or slope can be created to allow water to flow out of the drain. (**a**) Thermal equilibrium. (**b**) Small potential difference (linear regime). (**c**) Beyond pinch-off (saturation). In this condition the maximum capacity of the canal has been reached and any further increase in the height difference only results in a change in the height of the waterfall that now exists at the drain; in other words, the amount of water flow remains constant (After [2])

4.2.2 Refinements and Extensions to Long-Channel Theory

The long-channel MOSFET *I–V* behavior described thus far agrees well with MOSFETs made with channels >10 μm, but must be modified for shorter channels. This is especially relevant to today's state-of-the-art MOSFETs that have very short channels ($<<1$ μm), which we shall examine in more detail in Sect. 4.3.

4.2.2.1 Channel-Length Modulation

A smaller channel length L in a MOSFET is desirable because it increases the drain current for given bias conditions and also improves the ac and switching behavior. Thus, reducing L has been a design objective from the beginning of MOS transistor

design. However, with a shorter channel the drain current does not stay constant at I_{Dsat} as predicted by long-channel theory, but instead increases as the drain voltage exceeds V_{Dsat}. One important factor that contributes to this increased current is so-called channel-length modulation, which refers to the effective decrease in channel length as the drain voltage is increased beyond the pinch-off point in saturation (see Fig. 4.12c). In practice, MOSFET channel-length modulation is often treated analogously to the bipolar transistor Early effect we saw before by modifying Eq. (4.31):

$$I_{\mathrm{Dsat}} = \mu_n C_{\mathrm{ox}} \frac{W}{2L} (V_{\mathrm{G}} - V_{\mathrm{T}})^2 \left(1 + \frac{V_D}{V_A}\right) \tag{4.32}$$

where V_{A} is the MOSFET analogue of the BJT Early voltage described in Chap. 3.

4.2.2.2 Body-Bias Effect

Thus far we have considered the MOSFET under the condition that the source and substrate (or bulk) are held at the same voltage (i.e., $V_{\mathrm{S}} = V_{\mathrm{B}}$). We saw earlier for the MOS system that a voltage applied between the channel and the substrate caused the threshold voltage to increase. In a MOSFET this will result in a reduction of the drain current and is known as the body-bias or substrate-bias effect.[14] The effect of substrate bias can be qualitatively understood by considering two-dimensional band edge diagrams for the MOSFET structure as illustrated in Fig. 4.15. We can compare these types of band structures at zero substrate bias and at a negative substrate bias, for different MOSFET gate voltages, corresponding to flat band and inversion. In the presence of substrate bias the threshold voltage increases according to Eq. (4.17):

$$\Delta V_{\mathrm{T}} = \frac{\sqrt{2\varepsilon_s q N_{\mathrm{a}}}}{C_{\mathrm{ox}}} \left(\sqrt{2|\phi_{\mathrm{p}}| + V_{\mathrm{S}} - V_{\mathrm{B}}} - \sqrt{2|\phi_{\mathrm{p}}|} \right) \tag{4.33a}$$

and similarly using Eq. (4.19), for a *p*-channel device,

$$\Delta V_{\mathrm{T}} = -\frac{\sqrt{2\varepsilon_{\mathrm{s}} q N_{\mathrm{d}}}}{C_{\mathrm{ox}}} \left(\sqrt{2|\phi_{\mathrm{n}}| + V_{\mathrm{B}} - V_{\mathrm{S}}} - \sqrt{2|\phi_n|} \right) \tag{4.33b}$$

4.2.2.3 MOSFET Ion Implantation

Precisely controlling the MOSFET threshold voltage is critical in most integrated circuit applications. Ion implantation is a very important tool in IC technology that can be used to tailor the threshold voltage of MOSFETs on a wafer. Ion implantation allows precise control of surface dopant density and position. Typically the implanted atoms are used to create a region of increased dopant ion density in the

[14] Sometimes also referred to simply as *back-biasing*.

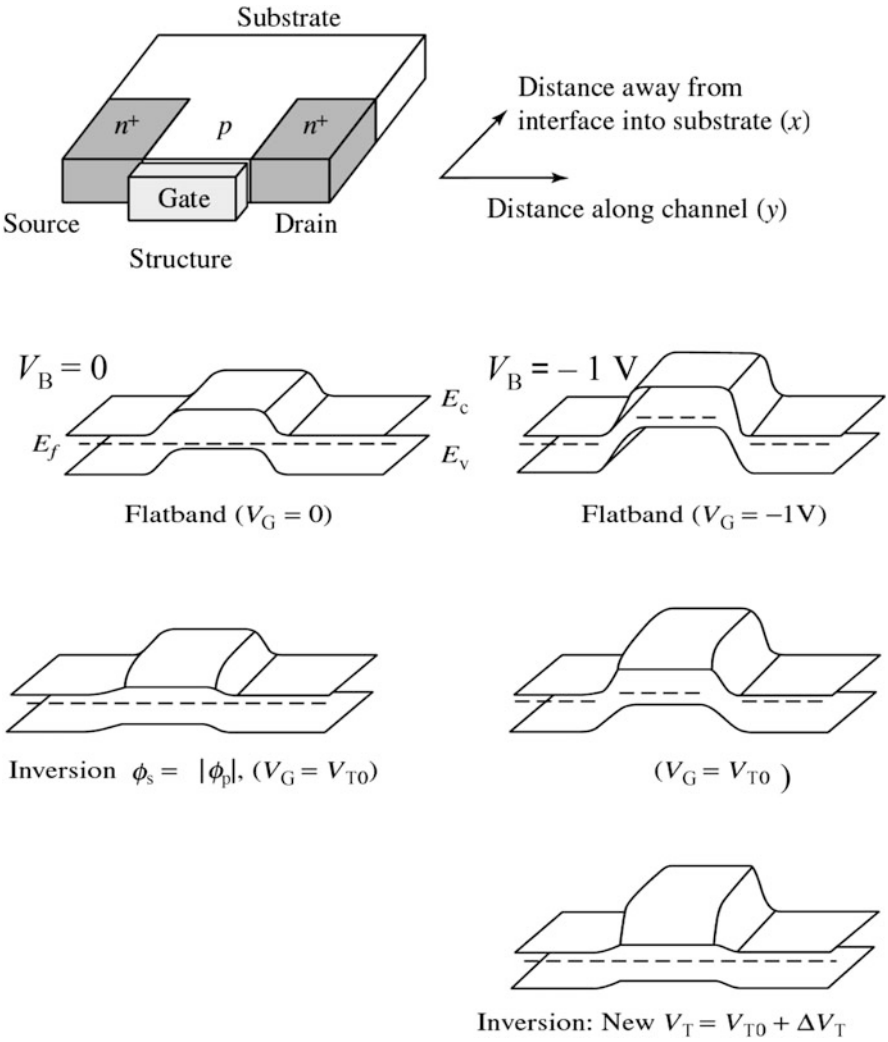


Fig. 4.15 MOSFET two-dimensional band edge diagrams illustrating the effect of substrate bias. The left-hand column illustrates the MOSFET without substrate bias. In the right-hand column a bias of −1 V is applied to the substrate, which shifts the flat-band voltage and results in a new (larger) threshold voltage in order to reach inversion and switch on the channel (Adapted from [2])

semiconductor near the oxide surface of the MOS system. This changes the threshold voltage predominately through the term

$$\Delta V_T = \frac{\pm q N_I}{C_{ox}} \tag{4.34}$$

where N_I is the number of dopant atoms per unit area implanted into silicon. Note the change will be positive if implanted with acceptor atoms and negative if implanted with donor atoms.

This technique is useful because often the threshold voltage without implantation is too close to zero for reliable fabrication of *n*-channel transistors. Thus ion implantation allows one to create commercially viable MOSFET circuits that are precisely controlled, in addition to allowing depletion- and enhancement-mode devices to be fabricated on the same IC, without the need to change material systems.

4.2.2.4 Bulk-Charge Effect

Our derivation of the drain current for the MOSFET that led to the long-channel equation was based on the assumption that the charge Q_d, due to ionized acceptors in the depletion region near the surface, is constant along the channel. However, we know that since the channel voltage increases from source to drain, the width of the depletion region also increases along the channel direction, y. Therefore the fixed space charge (sometimes called the *bulk charge*) will also increase closer to the drain as illustrated in Fig. 4.12b.

The variation in bulk charge along the channel means that the threshold voltage also depends on y, being greater near the drain. As a result, the bulk-charge effect causes drain current to be somewhat less than what simple long-channel theory predicts.[15] The modified long-channel drain current equation that takes this effect into account has the following form [6]:

$$I_D = \mu_n \frac{W}{L} \left\{ \begin{array}{l} C_{ox}\left(V_G - V_{FB} - 2|\phi_p| - \frac{1}{2}V_D\right)V_D \\ -\frac{2}{3}\sqrt{2\varepsilon_s q N_a}\left[\left(2|\phi_p| + V_D - V_B\right)^{3/2} - \left(2|\phi_p| + V_S - V_B\right)^{3/2}\right] \end{array} \right\} \tag{4.35a}$$

If $V_S = V_B = 0$ this reduces to

$$I_D = \mu_n \frac{W}{L} \left\{ \begin{array}{l} C_{ox}\left(V_G - V_{FB} - 2|\phi_p| - \frac{1}{2}V_D\right)V_D \\ -\frac{2}{3}\sqrt{2\varepsilon_s q N_a}\left[\left(2|\phi_p| + V_D\right)^{3/2} - \left(2|\phi_p|\right)^{3/2}\right] \end{array} \right\} \tag{4.35b}$$

[15] This is similar to the body-bias effect but with a changing (instead of constant) bias along the channel.

Instead of dealing directly with these equations the bulk-charge effect can also be accounted for by introducing a fitting parameter into the long-channel equations:

$$I_D = \mu_n C_{ox} \frac{W}{L} \left[\left(V_G - V_T - \frac{\alpha}{2} V_D \right) V_D \right] \tag{4.36a}$$

where α is called the *bulk-charge factor* and has a value greater than 1, typically around 1.5. The saturation voltage now becomes

$$V_{Dsat} = \frac{V_G - V_T}{\alpha}$$

and thus when the MOSFET enters the saturation region the current is now

$$I_{Dsat} = \mu_n C_{ox} \frac{W}{2\alpha L} (V_G - V_T)^2 \tag{4.36b}$$

4.2.3 Subthreshold Conduction

The first-order view of the MOSFET we have been pursuing as a device in which the gate voltage must reach V_T before any drain current can flow provides a useful picture for many applications. However, a small drain current flows in the MOSFET even when the gate voltage is below threshold and this can be very important for some applications. For example, we will see in Sect. 4.4 that a MOSFET is usually used to access the storage capacitor of a dynamic memory cell and even a small current flowing through the transistor allows the storage capacitor to discharge, destroying the stored information. The small drain current that flows when $V_G < V_T$ is called the *subthreshold current*.

In the subthreshold bias region, the drain voltage drops almost entirely across the reverse-biased drain-substrate depletion region. Thus the drift current component in the channel is negligible. However, at the source end of the MOSFET the applied gate voltage causes band bending at the surface (cf. Figs. 4.10 and 4.15) which reduces the size of the barrier to electron transport from the heavily doped source region to the channel, similar to a bipolar transistor, and this picture is illustrated with a band edge diagram in Fig. 4.16a: Electrons are injected from the source (which acts like a bipolar emitter region) into the p-type surface region (which acts like the base region of a BJT) and finally collected at the drain (like a BJT collector).

Thus, at gate voltages approximately 0.2 V below V_T the drain current is found to vary exponentially with gate voltage, and the subthreshold drain current can be written

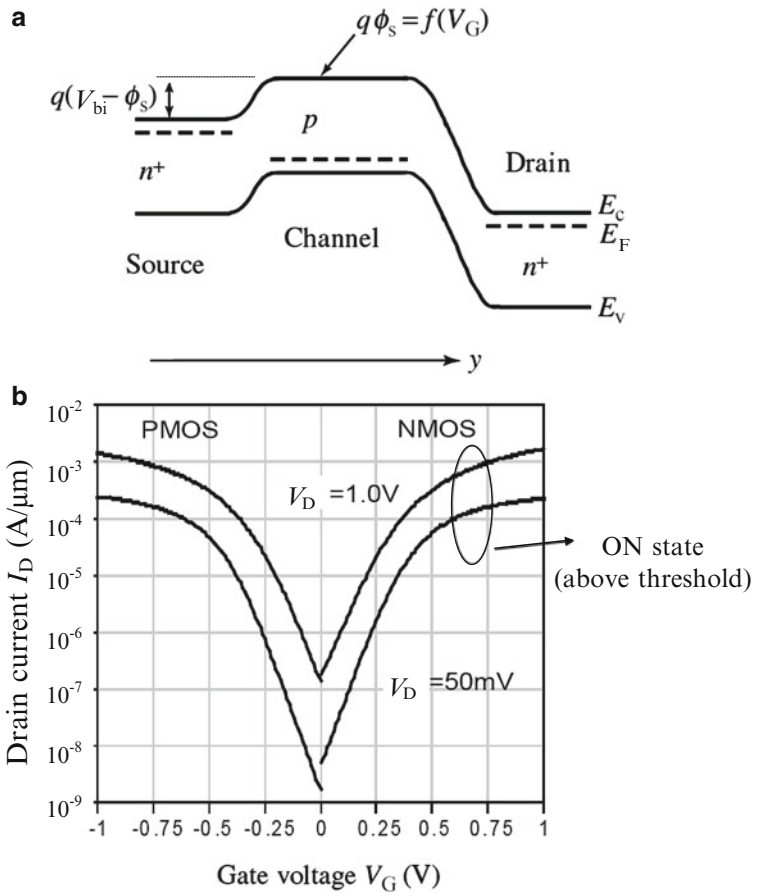


Fig. 4.16 (**a**) MOSFET subthreshold band edge diagram from source to drain. The applied gate voltage reduces the built-in potential barrier of the channel-source *pn* junction by modifying the surface potential, which leads to bipolar-like diffusion of carriers from source to drain below threshold. b. MOSFET subthreshold *I–V* data (current normalized per unit channel width) (After P. Packan et al., Intel Corp., 2009)

$$I_D \approx I_{D0} \exp\left(\frac{qV_G}{\eta k_B T}\right). \tag{4.37}$$

That is, the variation of subthreshold drain current with V_G is linear on a semi-log plot as shown in Fig. 4.16b. The slope of the linear region is important[16] because it determines the ratio of current flowing in the MOSFET between on/off states. Meeting a specific design requirement for this ratio places a lower limit on V_T for a given slope.

[16] Equation (4.37) tells us that the subthreshold current of a MOSFET can at most vary by one decade per 60 mV gate bias (often referred to as the *subthreshold slope*) at room temperature.

4.2.4 Small-Signal Parameters[17]

As with the bipolar transistor, the MOSFET transconductance g_m relates the output and input of the transistor, and in this case since the output (drain) current varies in response to a changing input (gate) voltage we have

$$g_m \equiv \frac{\partial I_D}{\partial V_G} \tag{4.38a}$$

For $V_D < V_{Dsat}$ the drain current was given by the long-channel equation (Eq. 4.29b), and therefore,

$$g_m = \mu_n C_{ox}(W/L)V_D \tag{4.38b}$$

i.e., g_m increases linearly with increasing drain voltage, but is independent of gate voltage in this region of operation. When $V_D > V_{Dsat}$ the transconductance is, to first order,

$$g_{msat} = \mu_n C_{ox}(W/L)(V_G - V_T) = \frac{2I_{Dsat}}{(V_G - V_T)} \tag{4.38c}$$

Thus in the saturation region the transconductance is independent of V_D, but depends linearly on V_G (which is useful for amplification). To obtain a simple low-frequency small-signal equivalent circuit one usually assumes infinite input resistance, while the small-signal output drain conductance is given by

$$g_d \equiv \frac{\partial I_D}{\partial V_D} \tag{4.39a}$$

This results in the equivalent circuit shown in Fig. 4.17a. Note that in saturation, to first order, g_d is zero, i.e., the small-signal channel resistance is infinite.

One intrinsic limit on the response time of a MOSFET is the channel *transit time*. For the long-channel MOSFET we can estimate the transit time starting from Eq. (4.29a):

$$\int_0^y I_D dy = \mu_n W C_{ox} \int_0^V [V_G - V_T - V(y)]dV$$

[17] As in previous chapters, we do not include non-idealities and other refinements when discussing the MOSFET small-signal parameters, but these can be included in a straightforward manner if necessary.

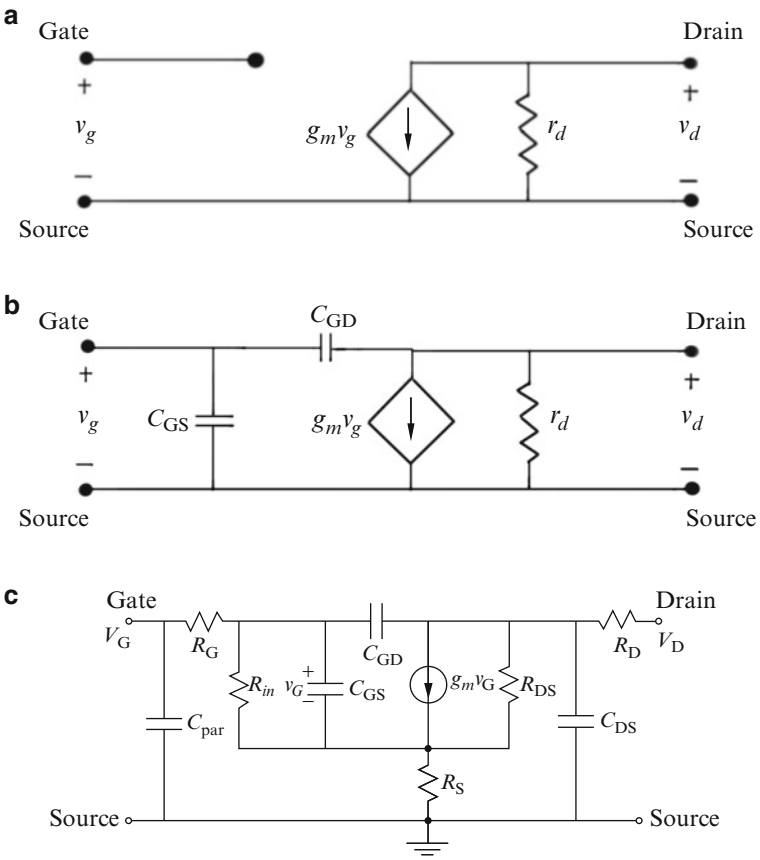


Fig. 4.17 MOSFET small-signal equivalent circuits. (**a**) First-order low-frequency circuit with infinite input impedance. (**b**) Equivalent circuit including gate-channel capacitances. (**c**) Equivalent circuit including parasitics. C_{GS} and C_{GD} are the "good" capacitances, while the remainder consists of parasitic coupling of the gate to contacts, contacts to substrate, various resistances, etc. (Adapted from [1])

Performing the integrations and isolating for $V(y)$ give

$$V(y) = (V_G - V_T) - \sqrt{(V_G - V_T)^2 - \frac{2I_D y}{\mu_n W C_{ox}}} \tag{4.40}$$

where the threshold voltage is taken to be independent of y, i.e., the bulk-charge effect discussed earlier is not considered. We can now find the electric field (in saturation):

$$E_y = -\frac{\partial V(y)}{\partial y} = -\frac{(V_G - V_T)}{2L} \frac{1}{\sqrt{1 - y/L}} \tag{4.41}$$

The transit time along the channel can be found using the electric field expression:

$$T_{\text{tr}} = \int_0^L \frac{1}{v_y}\, dy = -\int_0^L \frac{1}{\mu_n E_y}\, dy \tag{4.42a}$$

which finally gives

$$T_{\text{tr}} = \frac{4}{3}\frac{L^2}{\mu_n (V_{\text{G}} - V_{\text{T}})} \tag{4.42b}$$

Although this derivation is approximate and does not consider velocity saturation of the carriers in the channel, it gives a reasonable estimate of the transit time for most purposes.

The slowest response times in MOSFETs are not usually associated with the channel transit time, but with the time needed to charge capacitances associated with the device. For example, in the equivalent circuit of Fig. 4.17b, an approximate expression for the frequency at which the ratio of input to output current is unity, i.e., the cutoff frequency, is given by (below saturation)

$$f_{\text{T}} = \frac{\mu_n V_{\text{D}}}{2\pi L^2} \tag{4.43}$$

Similarly, in saturation the cutoff frequency also contains the L^{-2} dependence.[18] In practice, however, the speed of a MOSFET is almost always limited by the time required to charge the various parasitic capacitances of the device and elements it is connected to as illustrated in Fig. 4.17c. A decrease of 20–25 times in operating frequency is quite common for a modern high-density integrated circuit MOSFET compared to the intrinsic speed of the device.

4.3 Integrated Circuit Applications and MOSFET Scaling

4.3.1 Comparison of BJTs and MOSFETs

4.3.1.1 Input Resistance

We have seen that the dc input resistance is virtually infinite in a MOSFET. For a BJT the input resistance, while large, is much smaller than the MOSFET, mainly because of minority carrier injection into the forward-biased emitter that was discussed in Chap. 3.

[18] In other words, the cut-off frequency is inversely proportional to the transit time.

4.3.1.2 Transconductance

For a BJT we found the transconductance was equal to

$$g_m = \frac{I_0}{V_t}\exp\left(\frac{V_{BE}}{V_t}\right) = \frac{I_C}{V_t} \tag{4.44}$$

Thus the ratio of BJT to MOSFET (saturation) transconductance is

$$\frac{g_m(\text{BJT})}{g_m(\text{MOSFET})} = \frac{\frac{I_C}{V_t}}{\frac{2I_{Dsat}}{V_G}} \tag{4.45a}$$

where it has been assumed that $V_G >> V_T$. For equal values of output current, $I_C = I_{Dsat}$, and taking $V_G = +2.5$ V, this ratio becomes

$$\frac{g_m(\text{BJT})}{g_m(\text{MOSFET})} = \frac{2.5}{2 \times 0.026} \approx 50 \tag{4.45b}$$

i.e., the transconductance of BJT is typically much larger than that of a MOSFET.

The high transconductance of the BJT means it is capable of supplying higher current. In applications with appreciable load capacitance this means the BJT can charge and discharge the capacitance more quickly than a MOSFET, and thus BJTs are usually used for high-frequency circuits (e.g., 10 GHz and above) where their gain and power handling capabilities can be exploited. However, in the last four decades MOSFETs have displaced bipolar transistors to become the most extensively used active solid-state devices. The main reason for this is that MOSFETs have several advantages for high-density digital integrated circuits—an application area which has grown exponentially in the past 40 years: MOSFETs require fewer and simpler processing steps than BJTs and are thus much cheaper to manufacture. In addition it is easier to build MOSFETs in dense arrays as illustrated in Fig. 4.18. A BJT takes up roughly 5–10 times the area of a MOSFET on the surface of a wafer, and these differences are usually magnified further in circuit blocks containing several devices. Because of their simpler fabrication, higher density, and lower power dissipation, MOSFETs are widely used in memory and logic circuits, especially microprocessors.

4.3.2 MOSFET IC Applications

In this section we outline some of the major applications of MOSFETs that have resulted in VLSI and beyond.

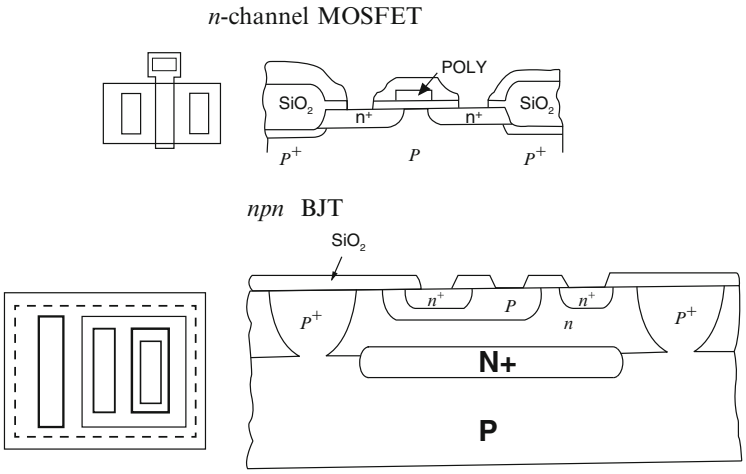


Fig. 4.18 Planar device area: MOSFET vs. BJT. The device cross sections and plan layouts illustrate the simpler and smaller planar MOSFET structure (Adapted from C. T. Sah, Proc. IEEE **76**, 1280 (1988))

4.3.2.1 Dynamic Random Access Memory (DRAM)

Since the MOS capacitor contains a variable amount of mobile charge it can be used to store information. As we have seen, the MOS capacitor can be biased into inversion or deep depletion depending on the frequency of the applied gate voltage. As shown in Fig. 4.19a, these two states of the MOS capacitor can represent two different memory states. Due to thermal carrier generation and leakage currents a MOS capacitor in deep depletion will eventually fill with mobile electrons and thus this type of memory is termed *dynamic*. MOS memory is organized in arrays to allow access to each cell in a *random* fashion (thus the term DRAM) (Fig. 4.19b).

The standard DRAM cell consists of 1 MOSFET, which serves as a switch to access one bit of information stored on 1 MOS capacitor (1T-1C DRAM cell; see schematic Fig. 4.19c). To understand the operation of a DRAM cell we need to examine the device structure. Figure 4.19c illustrates the basic structure of an individual IC DRAM cell:

When the word line is made higher than V_T the bit line is connected to the storage capacitor through the conducting MOSFET channel and can be used to write states into the capacitor (the capacitor plate is always held at a voltage V_{DD}, greater than the inversion threshold voltage). The bit line essentially changes the bias between the storage capacitor inversion layer and the underlying substrate. For 0V (logic '0') applied to the bit line the MOS capacitor becomes full of mobile inversion charge and when the word line (or pass transistor) is disconnected the charge remains and we have stored a "0" into the DRAM cell (this is the stable state of the capacitor). On the other hand if V_{DD} (logic '1') is applied to the bit line the inversion charge is

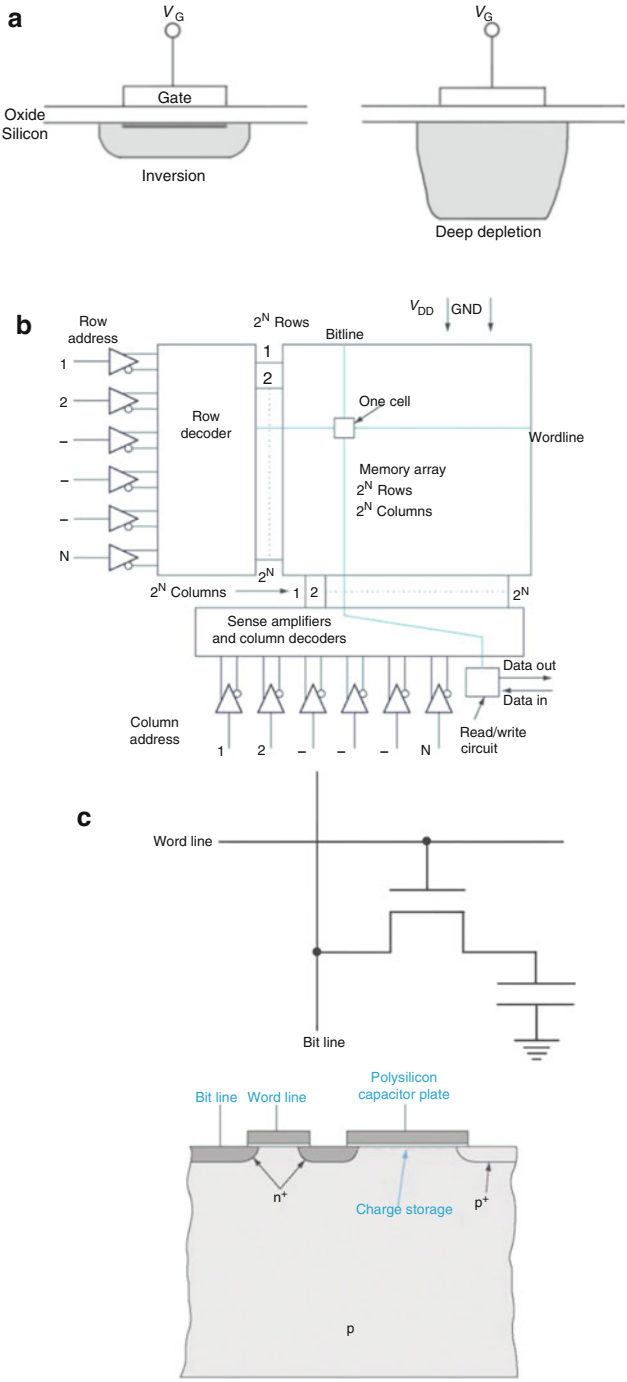


Fig. 4.19 MOS capacitor charge storage memory (DRAM). (**a**) Two charge states used to represent one bit of information. The MOS system is driven into either deep depletion (no mobile charges) or inversion (many mobile charges) (see Fig. 4.8). (**b**) Memory array layout for random access to a cell. (**c**) Basic structure of an individual 1T-1C DRAM cell (After B. Streetman, S. Banerjee, *Solid State Electronic Devices*, 6th Edition, Prentice-Hall, 2005)

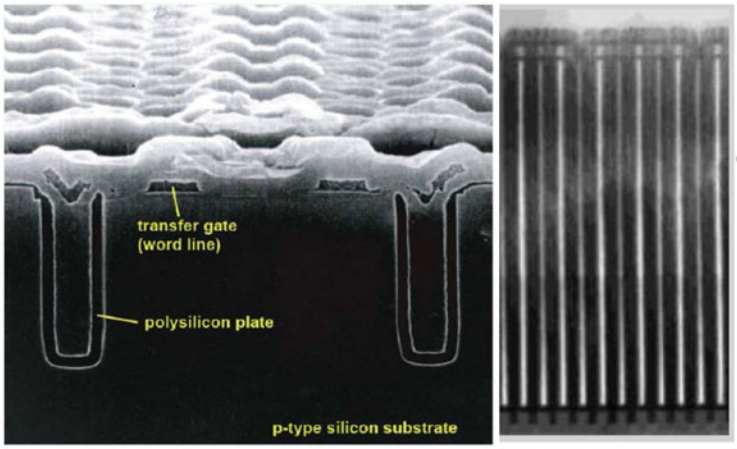


Fig. 4.20 Nonplanar DRAM IC structure cross sections displaying trench geometries to increase surface area. As memory density increases the aspect ratio (width/height) of the trenches must also increase to maintain adequate charge on the capacitor. Similar comments apply to the number of fins or layers in a stacked geometry (*Sources*: H. Sunami et al., IEDM Tech. Dig. 806, December 1982; S.- K. Park, Hynix Inc., 2011)

drawn out of the capacitor and it is driven into deep depletion. When the pass transistor is turned off we are left with an empty capacitor and a "1" has now been stored in the DRAM cell. Since MOS memory is dynamic, logic "1" will degrade toward logic "0" over time and therefore DRAM logic levels need to be refreshed periodically (~milliseconds).

To read the cell, the pass transistor is turned on and the stored charge is connected to the bit line, whose capacitance is altered depending on the amount of stored charge in the MOS capacitor. This leads to a voltage change on the bit line which can then be amplified, etc. to sense the state of the DRAM cell. After a read operation the charge state of the capacitor is destroyed and thus a read pulse has to be followed by a rewrite pulse to maintain the original stored information.

As memory density has continued to increase the area of each DRAM cell on an integrated circuit has in turn become smaller and smaller. Since the total capacitance (and thus the stored charge) of each MOS capacitor depends on its area, the simple planar MOS structure cannot be scaled down indefinitely without losing the ability to store adequate levels of charge that can be sensed by external circuitry. Starting with the 4 Mb generation (1980s) the required capacitor area could no longer be achieved with a planar geometry and thus the third dimension (*perpendicular* to the substrate) began to be used: Modern DRAM chips employ variations of either *trench* or *stacked* capacitor structures as illustrated by the images in Fig. 4.20.

4.3.2.2 Floating-Gate Memory

We have seen that DRAM is *volatile*, i.e., the information stored on the capacitors is lost after the supply voltage is cut. A very important type of *nonvolatile* memory is based on a MOSFET with *two gates*, one of which is floating. Figure 4.21a shows the device structure and schematic of a single floating-gate memory cell. Because the floating gate is completely isolated from any external circuitry it takes a very long time to reach thermal equilibrium and can retain excess charge almost indefinitely (i.e., a metastable state). If negative charge is stored on the floating gate the threshold voltage of the MOSFET will increase through the term:

$$\Delta V_{\mathrm{T}} = \frac{|Q_{\mathrm{fg}}|}{\varepsilon_{\mathrm{ox}}} d_1 \tag{4.46}$$

and thus we can sense two memory states on the floating-gate memory cell based on the presence or absence of stored charge by measuring the current through the floating-gate MOSFET (Fig. 4.21b).

Writing or programming the floating-gate memory cell typically involves the use of high-energy (or hot) electrons formed by applying large voltages to overcome the high barriers to electron flow presented by the oxide layer or via an electric field-assisted tunneling process known as Fowler–Nordheim tunneling. There are several ways to erase floating-gate memory, but the most useful for electronic devices also involves the tunneling of electrons out of the floating gate. The floating-gate writing and erasing processes are illustrated in Fig. 4.21c. Floating-gate memory in the form of high-density nonvolatile FLASH memory has become a very important technology and is rapidly growing and expanding its range of application (particularly in mobile devices).

4.3.2.3 Complementary MOS (CMOS)

Circuits consisting of both *p*- and *n*-channel MOSFETs are referred to as complementary MOS (CMOS).

CMOS Inverter

Figure 4.22a depicts the circuit schematic for a CMOS inverter: for small input voltage the p-MOSFET is ON and the n-MOSFET is OFF; output is V_{DD}. If the input voltage is increased the n-MOSFET turns ON and the p-MOSFET OFF; output is ground. Thus the voltage transfer characteristic has the form shown in Fig. 4.22b. CMOS is the dominant technology used in digital ICs since CMOS circuits[19] have

[19] General binary logic functions such as NAND and NOR gates, etc., can also be implemented in a straightforward manner using CMOS circuits. (F. M. Wanlass, C. T. Sah, IEEE Int. Solid-State Circuits Conf., Philadelphia, PA, Feb. 1963.)

very low standby (dc) power dissipation. This is due to the fact that under dc conditions one of the two FETs is always turned off and thus almost all the power dissipation in CMOS circuits takes place only during switching. The steep, well-defined voltage-transfer characteristic is also very desirable in digital designs.

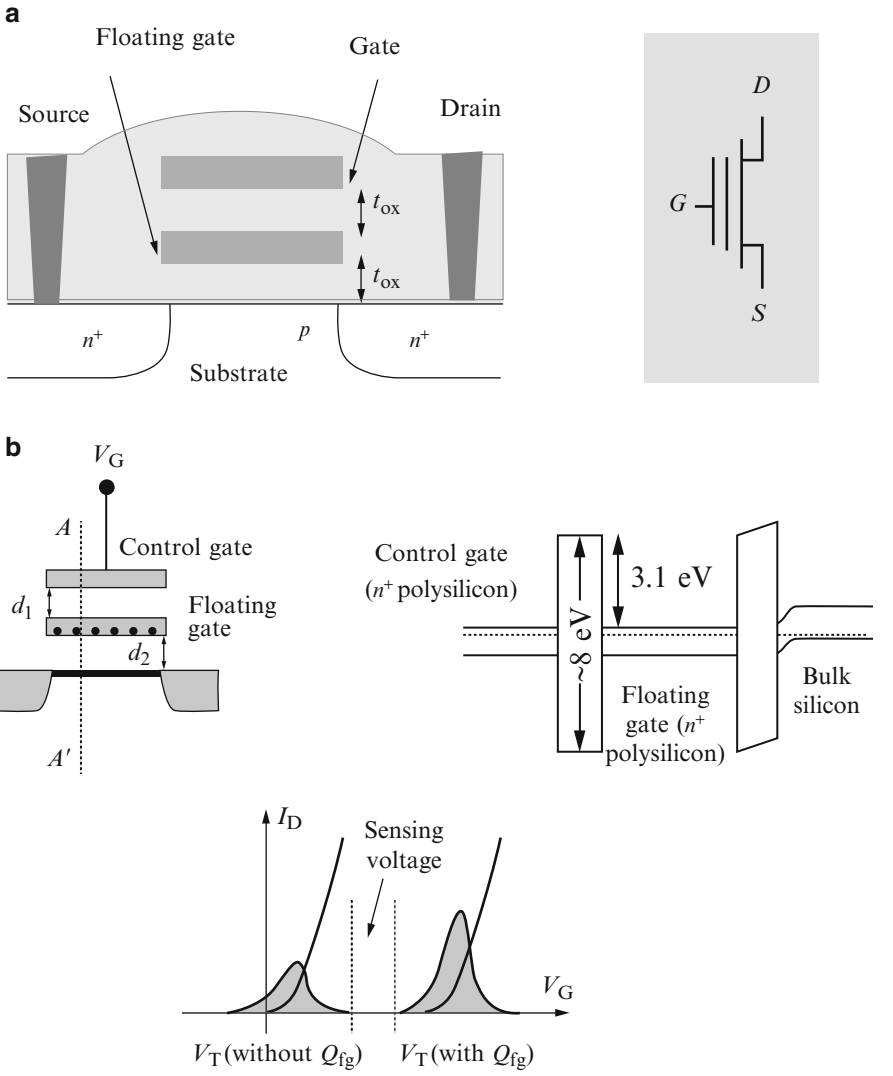


Fig. 4.21 Floating-gate nonvolatile memory. (**a**) Floating-gate MOSFET device structure and circuit symbol. The floating gate is electrically isolated and can store charge for several years. (**b**) Device cross section and band edge diagram. The MOSFET threshold voltage depends on the amount of charge stored on the polysilicon floating gate, which leads to different current values that represent the memory state of the device. (**c**) Writing and erasing floating-gate memory are accomplished by applying large voltages so that high-energy carriers can overcome the potential barrier (the field-assisted tunneling mechanism is illustrated) (Adapted from [2])

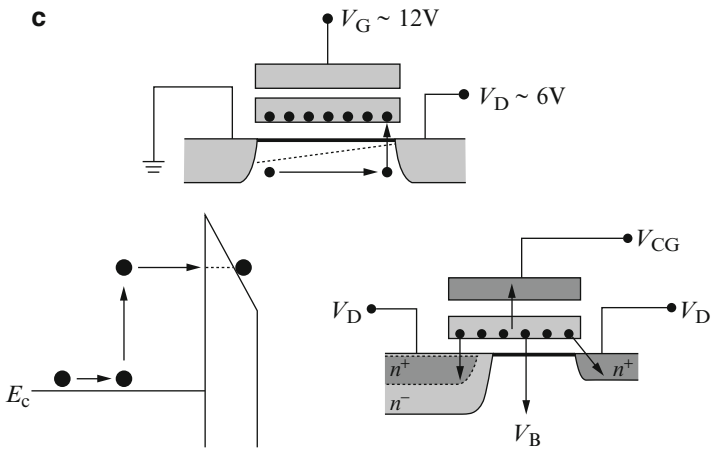


Fig. 4.21 (continued)

IC CMOS Structures

In order to fabricate integrated CMOS circuits on a single wafer we can either form p-wells on n-type wafers or n-wells on p-type wafers, which allows isolated *n*- and *p*-channel MOSFETs to be formed on the wafer surface.

In the design of a CMOS circuit manufacturing/fabrication process, an important consideration is the depth of the well region needed to avoid punchthrough from the bottom of the source/drain to the substrate below. If the junction depletion layers overlap large currents will flow and destroy the intended circuit function.

Example 4.2: CMOS Well-Depth Design Figure 4.23b shows the cross section of an n-well CMOS inverter. Given the n-well CMOS process data below, what is the minimum n-well depth that will avoid vertical punchthrough to the substrate?

$V_{DD} = 1.5\,\text{V}$
Substrate : $N_a = 5 \times 10^{14}\text{cm}^{-3}$
n-wells : $N_d = 3 \times 10^{15}\text{cm}^{-3}$
p-channel source/drain : $0.3\,\mu\text{m}$ depth; $N_a = 10^{18}\text{cm}^{-3}$

The analysis we used for the depletion width of a *pn* junction in Chap. 2 can be directly applied to find the necessary CMOS well depth. To find the required well depth for the given parameters we need the following equations from Chap. 2:

$$V_{bi} = \frac{k_B T}{q} \ln\left(\frac{N_d N_a}{n_i^2}\right)$$

$$x_d = x_n + x_p = \left[\frac{2\varepsilon_s}{q}\left(\frac{1}{N_a} + \frac{1}{N_d}\right)(V_{bi} - V_a)\right]^{1/2}$$

$$N_a x_p = N_d x_n$$

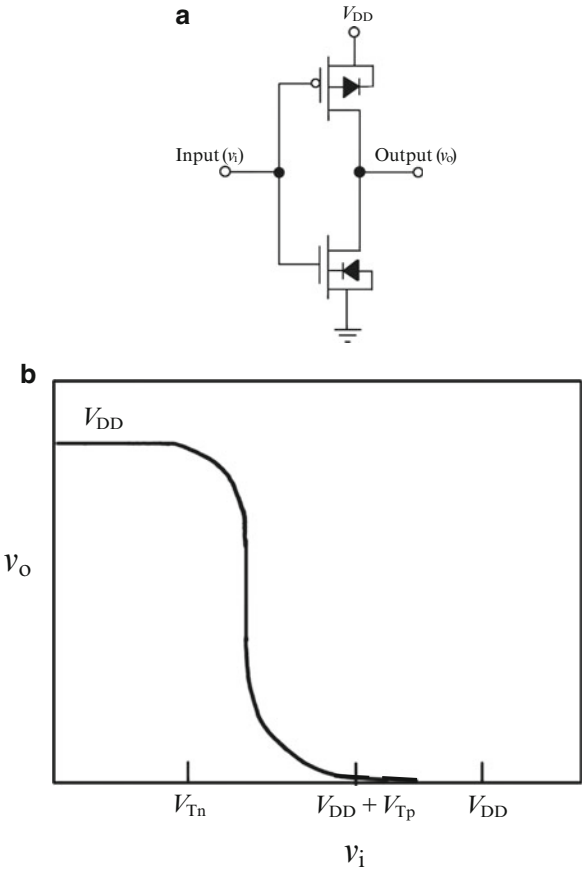


Fig. 4.22 (**a**) CMOS logic inverter circuit consisting of a *p*-channel (top) and *n*-channel (bottom) device with their channels connected in series with the supply voltage. (**b**) CMOS inverter voltage transfer characteristic showing relationship between input and output voltage (Note that the threshold voltage of the p-MOS device is negative)

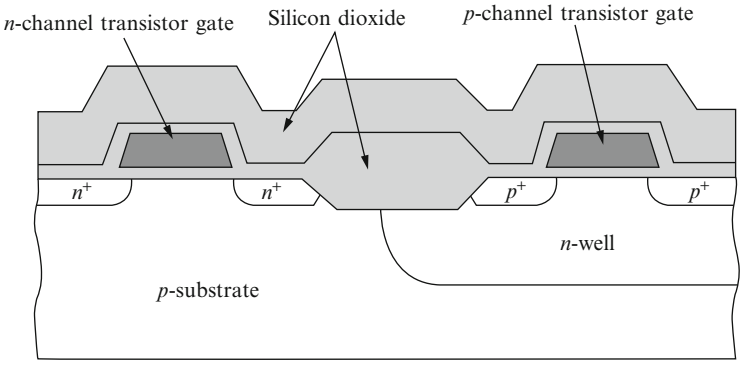


Fig. 4.23 CMOS IC structure showing n-well CMOS inverter cross-section. (Adapted from [2])

Vertical punchthrough occurs from the *p*-channel MOSFET source biased at V_{DD} to the grounded substrate. The source to n-well junction is essentially one sided and thus its depletion region extends primarily into the n-well. From the given data we can find $V_{bi} \sim 0.78$ V and the depletion width in the n-well is 0.58 μm. On the other hand, the *np* junction to the substrate has a built-in voltage $V_{bi} \sim 0.58$ V, and the total depletion width at 1.5 V reverse bias is 2.51 μm, of which 0.36 μm is in the n-well. The total depth of the n-well must therefore be at least

$$x_{\text{n-well}} = 0.3 + 0.58 + 0.36 = 1.24 \text{ μm}$$

The above analysis does not take into account that the drain of the *p*-channel MOSFET is at ground potential when it is turned OFF (see Fig. 4.22a). Thus the depletion region at the drain-well junction is wider than the source–well junction because of V_{DD}. However, even if the depletion regions between drain–well and well–substrate touch each other, no high punchthrough currents will flow because they are both at the same potential (ground). Despite this, it is generally not good design to have depletion regions touch because it may cause other regions of the well to become depleted of carriers or pinched-off and alter circuit operation. Performing a similar analysis to that above, the total minimum well depth needed to ensure that there are neutral regions throughout the well under all bias conditions is about 1.7 μm. Adding some margin for safety, a good design choice might make the well 2 μm deep.

4.3.3 MOSFET Scaling

The extremely high-volume production of MOSFET integrated circuits has led to an ongoing push to reduce device size, allowing more devices on a chip, and improving performance, all the while reducing cost. The minimum surface dimension in a MOSFET process is a key benchmark for the density of devices that can be built on a wafer. This dimension, typically taken to be the MOSFET channel length *L*, has continually decreased for over 40 years, and this trend is often described in terms of *Moore's Law*, i.e., the number of transistors in an IC will double every 18–24 months[20] (Fig. 4.24a). Over the past 40 years the minimum feature size has been reduced by a factor of about 500, and thus the area density of devices has increased more than 250,000 times (Fig. 4.24b). This enduring trend has led to more and more complex integrated circuits as illustrated by the chronology of images shown in Fig. 4.24c. Recently, the "More than Moore" concept shown in Fig. 4.24d has also become an emerging growth area: This approach typically combines non-digital functions (e.g., analog/RF, sensors, actuators, passive components, optoelectronics,

[20] The period for doubling has historically varied from about 1 to 3 years depending on the state of IC technology.

etc.) with traditional digital ICs into a single integrated package or chip in a way that complements traditional scaling to enable applications in new and diverse areas.

We saw earlier that the two major factors that control the intrinsic speed of MOSFETs are the channel (or gate) length and the speed (or mobility) of the channel carriers in traveling from source to drain. Both of these quantities, in addition to the gate capacitance, determine the drain current, which is of prime importance to IC

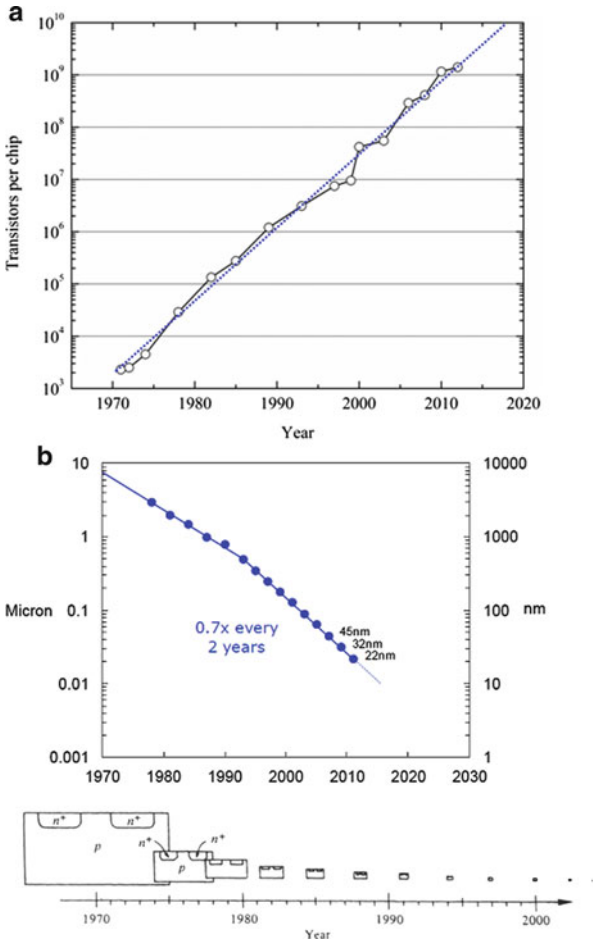


Fig. 4.24 Moore's Law. (**a**) Number of transistors per CPU versus year of production. Straight line fit (*dotted*) indicates approximately 100 times increase every 15 years. (**b**) Minimum feature size (nominal) versus year of production (a 0.7× (or $1/\sqrt{2}$) reduction in linear feature size results in a doubling of areal device density for each successive generation or node). (sketch adapted from [6]) (**c**) Chronology of integrated circuit chip optical images. (Data and images from Intel Corp.) (**d**) Illustration of "More than Moore" concept of functional diversification that incorporates multiple technologies (digital and non-digital) into an integrated system (*Source*: ITRS). Some of these concepts are discussed further in Chap. 5

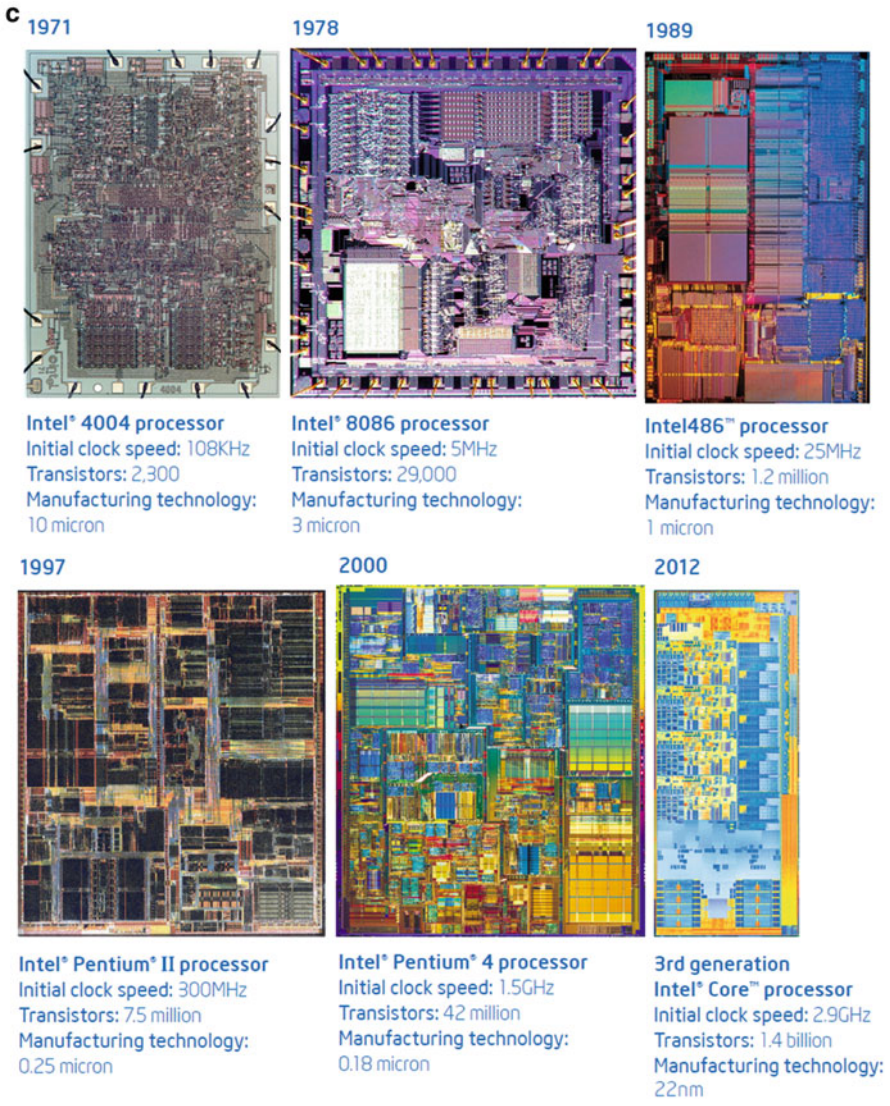


Fig. 4.24 (continued)

applications. In general, the main goals of scaling are (1) to reduce transistor size, (2) to increase drive current per unit width, and (3) to reduce power supply voltage.

As the gate length is reduced the other device parameters are scaled down concurrently in order for the transistor to function properly. So-called *scaling rules* have been devised as guides for this size reduction. In the simplest case, scaling rules ensure the internal electric fields of the MOSFET stay constant at reduced dimensions. As shown

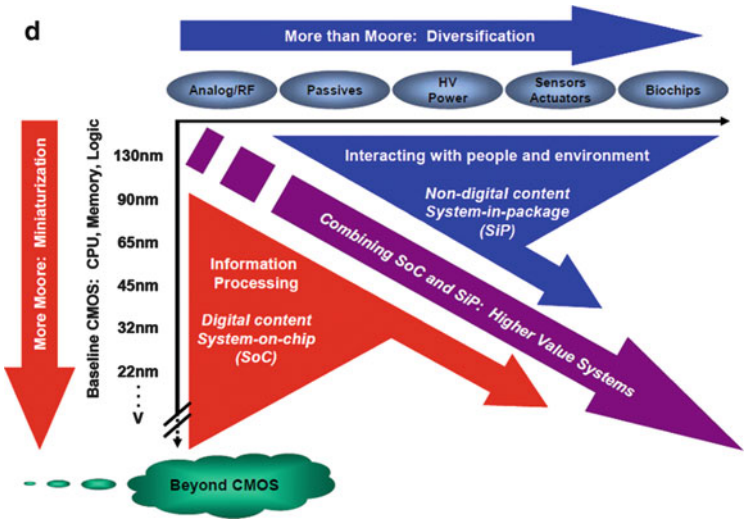


Fig. 4.24 (continued)

in Fig. 4.25, all device dimensions and voltages in this case are shrunk by a common scaling factor, whereas the doping level is increased by the same factor.[21]

In practice however, the direct use of any specific scaling rules is usually not possible. The main reason for this is that certain factors come into play at reduced channel dimensions that do not allow scaling in a straightforward manner, as described below.

4.3.3.1 Short-Channel Effects

1. Threshold Voltage Modification

The MOS threshold voltage needs modification in short-channel devices due to "charge sharing" between the source/drain depletion regions and the channel as shown in Fig. 4.26. As the source and drain are brought closer together their space-charge regions contribute a larger fraction of the MOS capacitor charge, which causes the threshold voltage to decrease relative to a long-channel device.[22] The above effect can be minimized by, for example, increasing the substrate doping level in order to decrease the source/drain *pn* junction depletion widths (see Chap. 2).

[21] The increased doping level is necessary to scale down the depletion width at the drain–substrate junction.

[22] In other words the MOS capacitor is partially charged by the depletion widths of the source and drain regions as they become comparable to the channel length and thus less charge needs to be induced by the gate voltage to reach inversion.

Device or Circuit Parameter	Scaling Factor
Device dimension t_{ox}, L, W	1/K
Doping concentration Na	K
Voltage V	1/K
Current I	1/K
Capacitance $\varepsilon A/t$	1/K
Delay time/circuit VC/I	1/K
Power dissipation/circuit VI	$1/K^2$
Power density VI/A	1

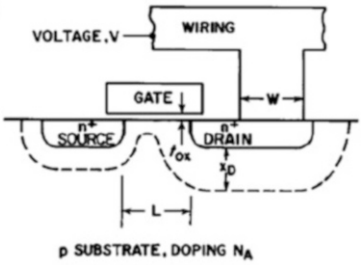


Fig. 4.25 Constant-field MOSFET scaling rules (After R. H. Dennard et al., IEEE J. Solid State Circuits **SC-9**, 256 (1974))

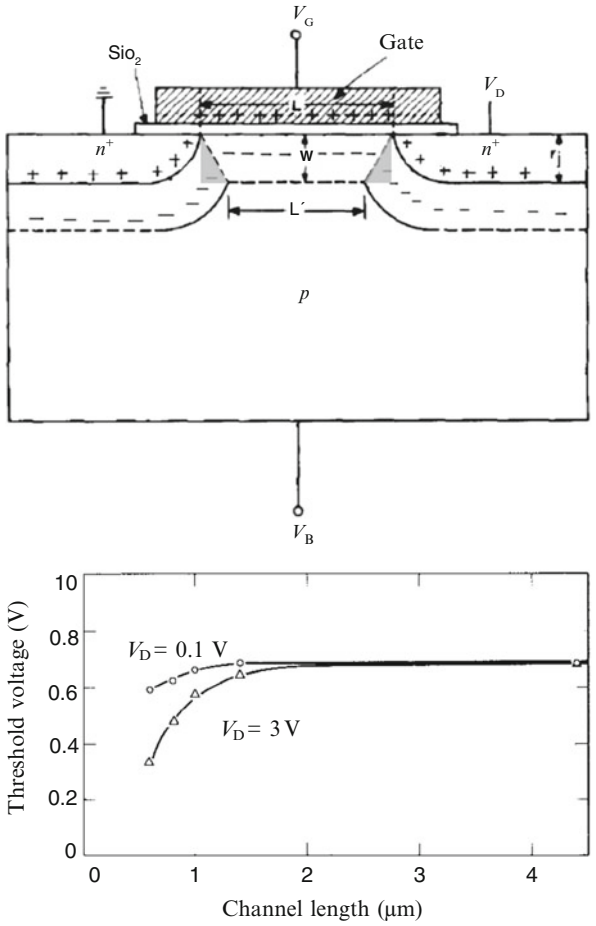


Fig. 4.26 MOSFET charge sharing in short-channel devices. The source/drain depletion regions provide a significant amount of the spacecharge beneath the channel as L decreases (shaded regions), which reduces the amount of charge that must be induced by the gate voltage and hence the threshold voltage becomes smaller (Adapted from L. D. Yau, Solid-State Electron. **17**, 1059 (1974).) This effect increases for larger drain voltages as shown by the experimental data (After Y. Taur et al., IEEE Trans. Electron Dev. **ED-32**, 203 (1985))

However, there is a limit to how much the doping level can be increased before junction breakdown becomes a problem.

On the other hand, channels that are very narrow (small W) will experience an *increased* threshold voltage due to the additional "fringing" field lines at the edges of the channel. In effect, the gate voltage has to induce additional charge in order to reach inversion (not accounted for by the ideal parallel plate capacitor treatment), which results in the increased threshold voltage value compared to wider channel devices.

2. Velocity Saturation and Mobility Degradation

In the short channels of aggressively scaled MOSFETs *velocity saturation* of carriers is more likely to occur due to the smaller distance over which the source–drain voltage changes: At very high electric fields the channel carrier drift velocity approaches a saturation value and thus the mobility decreases from its low-field value.[23] The reduction in mobility in turn places a cap on the maximum drain current of the MOSFET (Fig. 4.27a) and thus limits the current increase that can be attained by scaling compared to a device that obeys long-channel theory.

Mobility degradation is another factor that limits the performance of short-channel devices. This is caused by increased surface scattering due to the vertical or transverse electrical field induced by the gate voltage as illustrated in Fig. 4.27b.[24] Gate oxide fields are more important in short-channel transistors because the power supply voltage has not been scaled as aggressively as the gate oxide thickness in order to improve MOSFET performance (i.e., by increasing the channel inversion charge and hence the drain current).

3. Leakage Currents

Another important effect of scaling is related to the increased current that flows through the MOSFET in its off state. In addition to the subthreshold current discussed earlier increasing for shorter channels, there are several other mechanisms that can contribute to the off-state leakage current in a MOSFET, some of which are illustrated in Fig. 4.28a.

When the gate oxide becomes very thin, direct tunneling of electrons through it causes significant gate current to flow and thus reduces the input impedance. Conventional gate oxide thicknesses measure only a few atomic layers and cannot shrink much further without the effect of tunneling degrading performance beyond acceptable levels. The reverse-biased drain–substrate junction can also contribute to the leakage current via either impact ionization processes (avalanching) or band-to-band (Zener) tunneling (see Chap. 2).

[23] See Appendix A, Sect. A.2 for further details.

[24] See Appendix B for inversion layer mobility data as a function of transverse electric field.

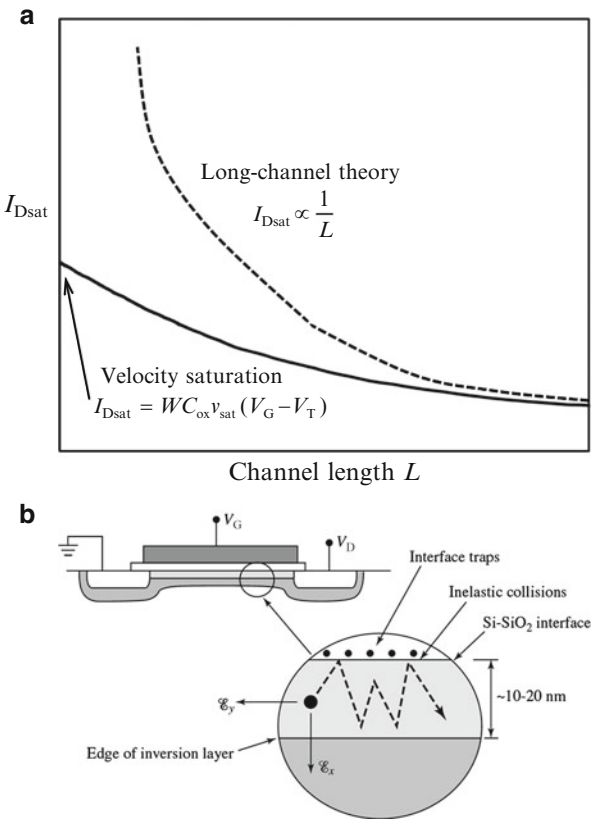


Fig. 4.27 (**a**) Effect of MOSFET channel carrier velocity saturation on drain current compared to a device obeying long-channel theory. (**b**) Illustration of mobility degradation caused by transverse electric field (After [2])

Furthermore, *punchthrough* from source to drain is another potential problem for short-channel MOSFETs that causes unwanted current flow. This can occur along the channel surface interface in a manner similar to what was discussed in Chap. 3 in relation to bipolar transistors (cf. Fig. 4.16a) and is usually referred to as *drain-induced barrier lowering*. In addition, *subsurface punchthrough* (Fig. 4.28b) can also occur as the source and drain depletion regions approach each other for short channels. Once again these effects can be suppressed by increasing the substrate doping level.

Minimizing the various leakage current contributions in order to constrain power dissipation is critical for the production of modern high-density integrated circuits.

4.3.3.2 MOSFET Scaling Below 100 nm

In order to meet the challenges of continued device scaling while maintaining performance, reliability, and cost-effectiveness, several important advancements to the basic silicon MOSFET structure have been introduced since sub-100-nm gate length transistors were first introduced commercially beginning in 2000. Some of

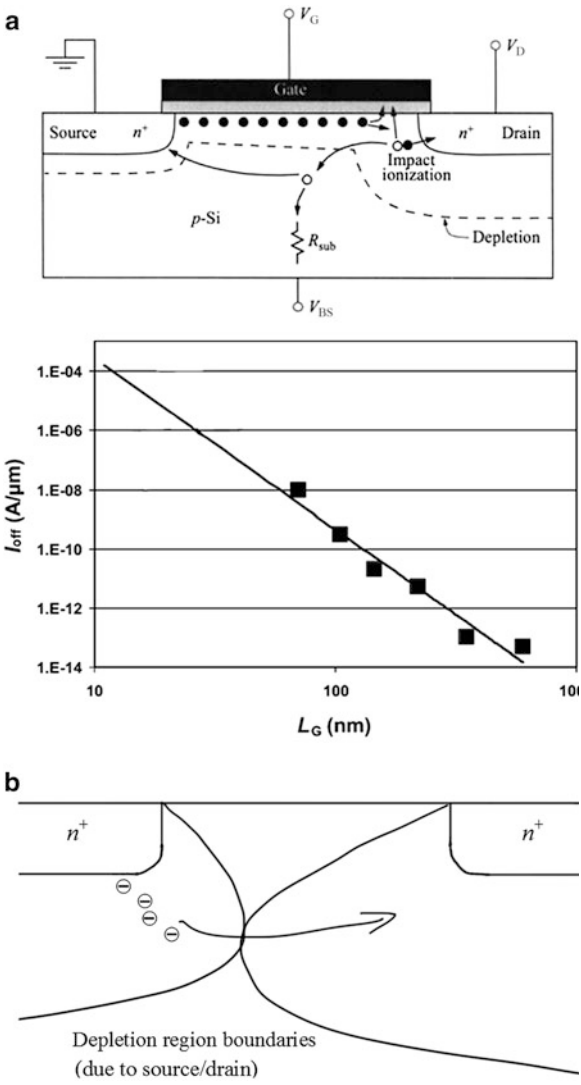


Fig. 4.28 (**a**) MOSFET "off"-state leakage current contributions. (After [1].) The data display off-state current vs. physical gate length of MOSFETs in commercial integrated circuits. (**b**) Illustration of subsurface source–drain punchthrough

the major changes in state-of-the-art MOSFET technology that have allowed scaling well below 100 nm are highlighted below.

1. Silicon on Insulator

One way to improve performance is by using silicon-on-insulator (SOI) structures. Here, MOSFETs are built directly on top of an oxide layer (e.g., SiO_2). SOI reduces leakage current and also decreases unwanted parasitic capacitance with the substrate as illustrated by the schematic in Fig. 4.29a.

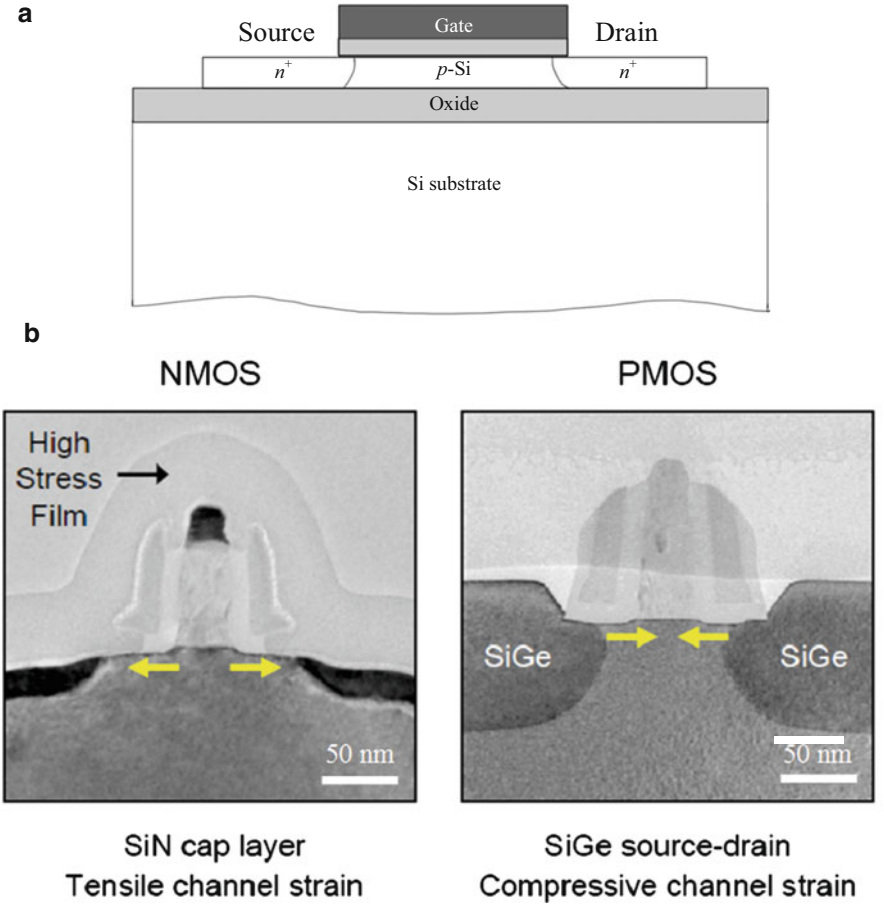


Fig. 4.29 (**a**) SOI MOSFET device schematic. By placing the MOSFET on an insulating layer, the drain junction area, and hence leakage current, is significantly reduced. In addition, parasitic capacitance with the Si substrate is reduced due to the separation introduced by the intervening oxide layer. (**b**) Strained MOSFET channels first used in the 90 nm node to increase Si channel carrier mobility (*Source*: Intel Corp.)

2. Strained Channels

A more recent innovation, used by Intel beginning with their 90-nm technology process (~50 nm gate length[25]), is the use of *strained* silicon for the channel, which can increase carrier mobility and thus increase the output drive current of the MOSFET relative to unstrained silicon. The increase in mobility is caused by the

[25] In several technology "nodes" etching is used to reduce certain device dimensions below the nominal or "printed" minimum feature size.

change in band structure (and thus effective mass) when the silicon lattice is stretched or compressed. One way to introduce strain into the channel is to epitaxially[26] grow a strained silicon layer on top of a SiGe alloy layer. Strain is introduced due to lattice mismatch between the layers. However, it is also possible to modify the strain in the channel by other means: The two examples shown in Fig. 4.29b depict how either a "capping" layer on top of the MOSFET or the addition of SiGe into the source and drain regions can create tensile or compressive strain into the channels of P- or N-MOS devices, respectively. Using this type of nanometer-scale "strain engineering" has allowed over 20–30 % increases in device performance and currently all major semiconductor electronics manufacturers employ some type of strain into their integrated circuit fabrication processes.

3. "High-κ" Gate Oxide

The 45 nm process was introduced in 2007 by Intel. A landmark change was that for the first time, the SiO_2 gate oxide was successfully replaced by a new high-κ (i.e., a high dielectric constant or relative permittivity compared to SiO_2) hafnium (Hf)-based dielectric material, which also included the use of metals for the gate material[27] (similar process developed at IBM) as shown in Fig. 4.30. High-κ materials allow the capacitance of the gate oxide to be increased and therefore increase output drive current and/or a reduction in gate leakage current by using a thicker gate dielectric, while still maintaining performance.

4. Three-Dimensional Channel Geometry

At the 22-nm level (2012) silicon MOSFETs with fundamentally different structures have been realized commercially, i.e., the *tri-gate* or "*3D*" FET (Fig. 4.31). This transistor has an almost *fully depleted body* (i.e., the space-charge region fills the entire channel), which leads to much lower leakage currents. In addition, by employing a nonplanar geometry the effective channel width increases and thus provides more current for the same footprint compared to a planar device[28] and also improves the electrostatic gate–channel coupling resulting in sharp subthreshold characteristics.

Lastly, we note that the innovations described above in 1–4 are often used in combination in order to achieve multiple benefits in modern integrated circuit electronics.

[26] See Appendix A, Sect. A.3.

[27] A metal gate reduces resistance and maximizes the capacitive coupling of the gate to the channel.

[28] This implies that a channel with a large aspect ratio (tall and thin) would in principle provide opportunities for future scaling similar to the nonplanar capacitors used for DRAM.

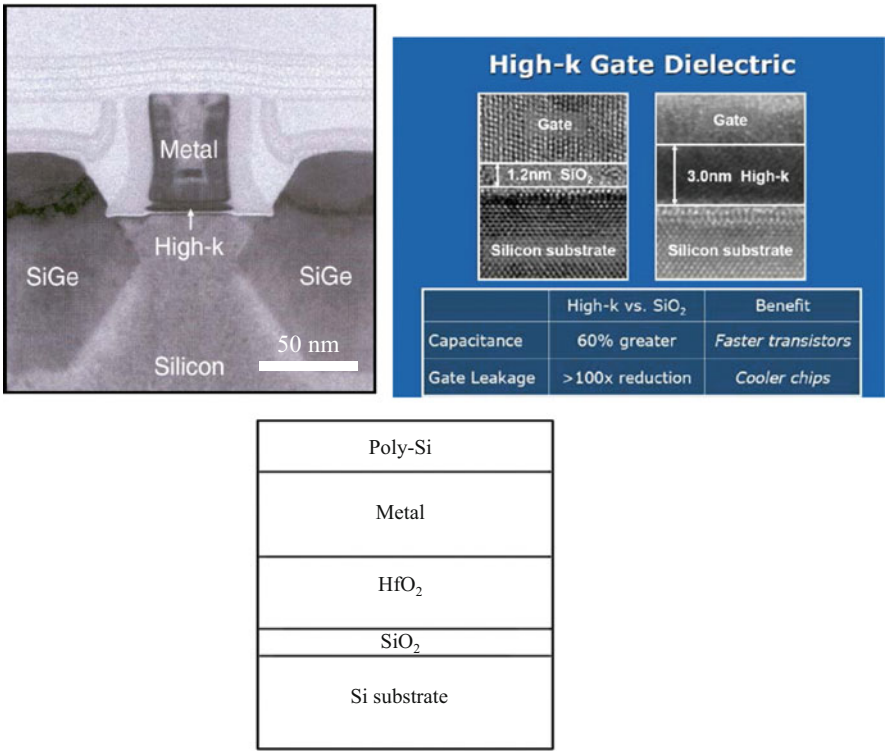


Fig. 4.30 High-κ dielectric and metal gate MOSFETs introduced at the 45 nm node (shown here for a *p*-channel device). (*Source*: Intel Corp.) A schematic high-κ stack layer structure is also shown

4.3.3.3 Additional Challenges for Continued MOSFET Scaling

Device performance and fabrication limits are linked to the decreasing ratio of feature sizes to fundamental dimensions (e.g., atomic, patterning wavelength) as scaling continues. At some point the ultimate limit of MOSFET scaling will be reached for very small gate lengths in one form or another, for example:

- Direct tunneling from source to drain
- Statistical variation of doping levels
- Increased parasitic capacitance
- Increased source/drain contact resistance
- Problems with strain engineering

Nevertheless, even more radical changes to silicon integrated circuit devices/processing[29] are being examined and developed, which could extend MOSFET

[29] We have focused predominantly on device-related challenges associated with scaling. However, there is a range of other important issues that must be considered such as interconnects, fabrication, process integration, etc. [5].

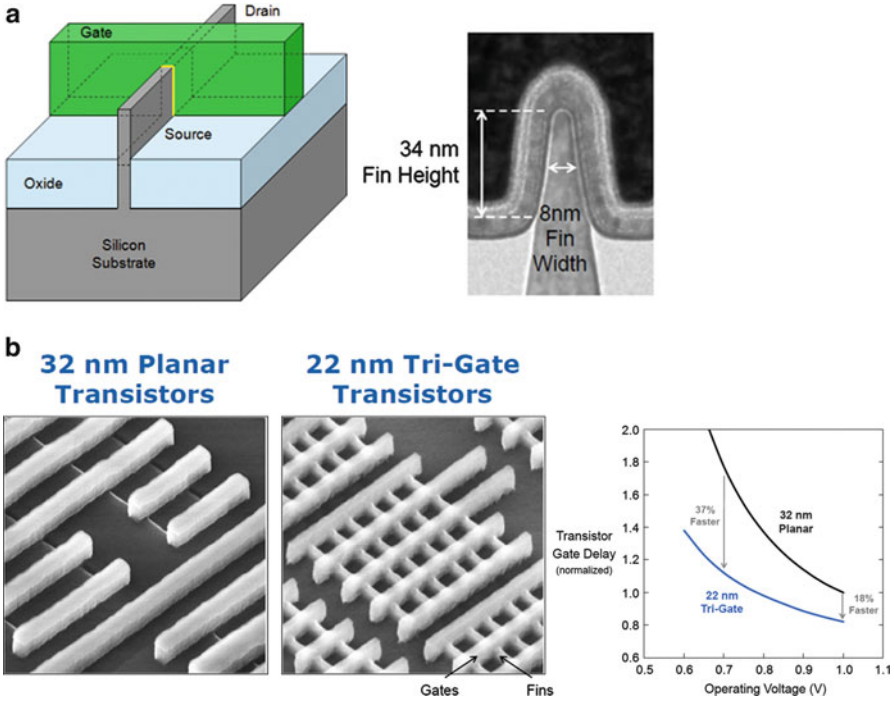


Fig. 4.31 Tri-gate FETs (Intel Corp.) (**a**) Device structure schematic. The *yellow region* indicates the gate oxide that wraps around the three sides of the channel fin. An electron microscope image shows a close-up view of the channel in cross section. (**b**) Structure and performance comparison of 32 nm planar MOSFETs to 22 nm tri-gate MOSFETs (Note that for the tri-gate structure multiple fins are connected together in order to increase output current)

scaling and Moore's Law for many years to come. In the next chapter (Sect. 5.1) we examine some of these possibilities.

References

1. Sze, S.M., Ng, K.K.: Physics of Semiconductor Devices, 3rd edn. Wiley Interscience, Hoboken (2007)
2. Muller, R.S., Kamins, T.I.: Device Electronics for Integrated Circuits, 3rd edn. Wiley, New York (2003)
3. Ng, K.K.: Complete Guide to Semiconductor Devices, 2nd edn. Wiley Interscience, New York (2002)
4. Streetman, B., Banerjee, S. *Solid State Electronic Devices*, 5th edn. Prentice-Hall (1999)
5. *International Technology Roadmap for Semiconductors*, ITRS, 2011–12; www.itrs.net
6. Pierret, R.F. *Field Effect Devices*, 2nd edn. Prentice-Hall (1990)

Problems

1. *MOS band edge diagrams*. Sketch the thermal equilibrium band edge diagram for a MOS system made with a gate material whose work function is greater than that of an n-type semiconductor substrate. (Assume zero oxide charge.)
2. *MOSFET design*[30]. A MOS system has a nickel gate (work function 5 eV) and 0.2 Ω-cm *p*-type silicon substrate. Assume there is an oxide surface charge density $(Q_f/q) = 3 \times 10^{10}\text{cm}^{-2}$. (1) If the oxide thickness is 5 nm, determine V_T (assume there is no additional substrate bias). (2) Design the channel dimensions so that the resulting MOSFET outputs a current of 1 mA when V_G is 1 V above threshold and $V_D = 0.75$ V. Assume long-channel theory applies with $\mu_n = 320\ \text{cm}^2\text{V}^{-1}\text{s}^{-1}$ and a bulk-charge factor of 1.4. (3) How small can your design be scaled before channel carrier velocity saturation at the source becomes a problem? (Hint: consider Eq. (4.25))
3. *Short-channel MOSFETs*. Several of the short-channel effects caused by MOSFET scaling can be controlled by increasing the substrate doping level. What limits how heavily the substrate can be doped?
4. *Limits of MOSFET scaling*. Consider direct source–drain tunneling as the factor that will ultimately limit silicon MOSFET scaling in the future to come up with a lower limit on channel length, L, when the end of scaling has been reached.

[30] This problem may be somewhat difficult and/or lengthy.

Chapter 5
Emerging Devices for Electronics and Beyond

"We keep moving forward, opening new doors, and doing new things, because we're curious and curiosity keeps leading us down new paths."

W. E. Disney

As evidenced in the preceding chapter, electronics continues to evolve. Indeed, as described in the introduction of Chap. 1 the field has been evolving and changing since its inception: From the first electromechanical/chemical devices in the early nineteenth century to the introduction of vacuum tube devices at the beginning of the twentieth century, and on to modern solid-state electronics and the integrated circuit of today. As the integrated circuit advances toward its final level of maturity, electronics appears to be approaching yet another important crossroad in the first part of the twenty-first century. What lies beyond may be as drastic as the change, for example, in going from vacuum tubes to solid-state devices. Although it is difficult to predict the precise path electronics will take, the tremendous success of silicon planar processing is also enabling emerging devices and applications in diverse areas beyond traditional electronics that are already beginning to make an impact.

This chapter discusses aspects of solid-state electronics moving forward in the near term (toward 2020/25) and some of the other emerging applications and extensions of this technology that are being developed.

5.1 Nanoelectronics

Electronics based on silicon MOSFETs has been scaled to nanoscale[1] dimensions, as discussed in Chap. 4. As scaling reduces device dimensions even further, silicon nanoelectronics will continue to be refined along with other materials to create deep

[1] The most common definition of the nanoscale is 1–100 nm. In this book, any device structure that has nanoscale extent in one or more spatial dimensions is considered a nanostructure.

C. Papadopoulos, *Solid-State Electronic Devices: An Introduction*,
Undergraduate Lecture Notes in Physics, DOI 10.1007/978-1-4614-8836-1_5,

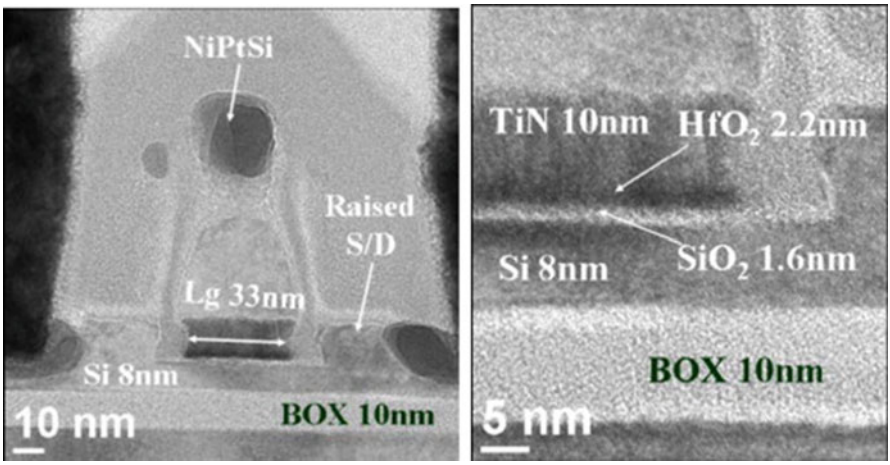


Fig. 5.1 Thin silicon channel used for FD-SOI structures. In this example an 8 nm layer of Si is used on top of a 10 nm buried oxide layer (After C. Fenouillet-Beranger *et al*., Solid-State Electron. **54**, 849 (2010))

nanoscale ICs over the next 10–15 years. In parallel with these developments, other types of nanoscale devices based on novel/different operating principles that provide potential advantages "beyond CMOS" (Si-based or otherwise) are beginning to appear. We cover each of these two broad areas in turn below.

5.1.1 Continued (MOSFET) Scaling[2]

Both planar and multi-gate MOSFET structures are being developed further as device dimensions approach the 10 nm level:

5.1.1.1 Si Nanoelectronics

In terms of silicon-on-insulator (SOI) devices, an additional enhancement is so-called *fully depleted* SOI (FD-SOI). Here, a very thin (~5–10 nm) Si layer is used for the channel as shown in Fig. 5.1. In FD-SOI the silicon channel is fully depleted of mobile carriers, which improves device performance compared to traditional or partially depleted SOI structures by further reducing leakage current and improving subthreshold slope. FD-SOI also eliminates the so-called floating-

[2] Although the focus in this section is on digital logic devices and circuits, much of the discussion also applies to RF/analog and mixed-signal applications, with appropriate modifications depending on the specific device functions required.

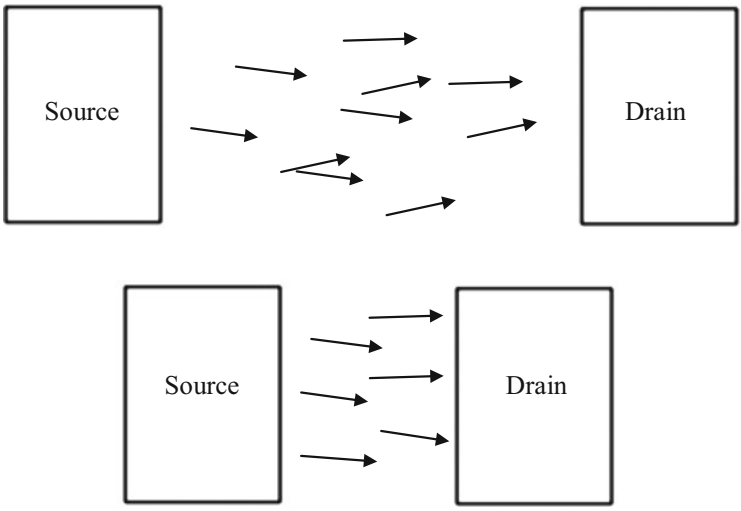


Fig. 5.2 Illustration of quasi-ballistic FET transport. In the *top diagram* carriers are scattered as they drift from source to drain. The *bottom diagram* illustrates the quasi-ballistic case where the channel length becomes comparable to the mean free path and thus carrier scattering is reduced leading to increased current flow in a typical MOSFET channel. In the limit of very small channel length the carriers do not experience any scattering and the transport is said to be ballistic within the channel (scattering may still be introduced by the source/drain contacts themselves)

body effect that can result in charge buildup and erratic device behavior. Fully depleted multi-gate (e.g., tri-gate) structures on SOI are a further evolution of this technology.

Continued improvement of Si MOSFET performance will require even higher κ dielectric materials ($\kappa > 30$) for the metal/gate stack, in addition to continued progress in strain-enhanced channels for both planar and nonplanar geometries. As the channel length is scaled below 20 nm it also becomes comparable to the mean free path of carriers traveling from source to drain (see Fig. 5.2). In this so-called quasi-ballistic regime the amount of carrier scattering is progressively reduced at smaller dimensions, which results in increased current compared to longer channel MOSFETs. Estimates predict that the ballistic transport current enhancement will reach a factor of about 2 for very short channel devices (~10 nm).[3]

5.1.1.2 Alternate Channel Materials

Once Si nanoelectronics has matured to the point where further improvement via scaling is no longer feasible, a more "radical" approach to improve MOSFET performance would be to use a completely different semiconductor altogether for

[3] ITRS 2012 update.

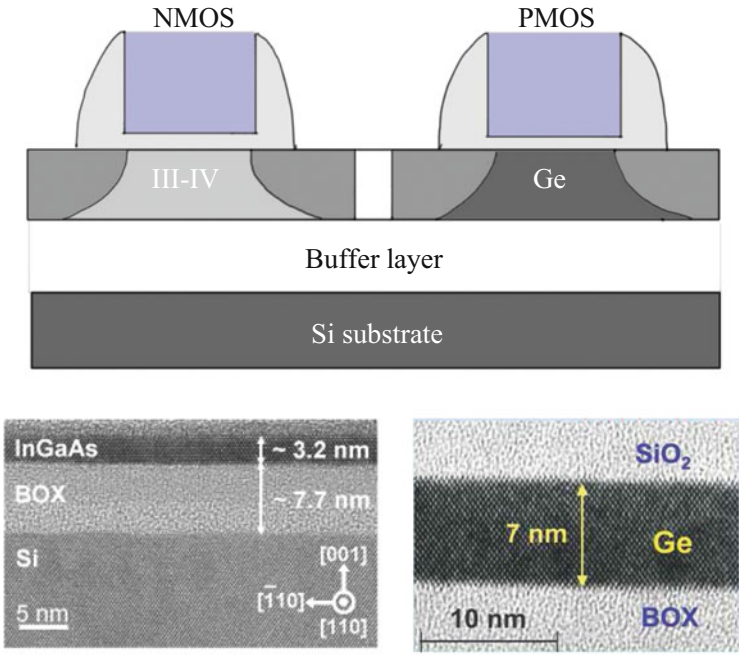


Fig. 5.3 III–IV and Ge MOSFETs integration schematic and channel cross sections of device structures (After: M. Yokoyama et al., VLSI Symposium, 12B-3, Kyoto, June 2009.) Both planar and multi-gate FET structures are possible using alternate channel materials

the channel itself. This would follow on the earlier paradigm set by the replacement of the traditional SiO_2 gate oxide with high-κ materials. In particular, several other semiconductors possess higher carrier mobilities than Si and thus allow performance improvements that would not otherwise be possible. For example, III–IV compounds have electron mobilities many times greater than Si. Similarly, Ge has a significantly larger hole mobility than Si.

The challenge for alternate channel materials is finding effective ways to integrate them into reliable large-scale circuits in a manner that is compatible with and can take advantage of existing silicon processing technology. Significant progress has been made using so-called heterointegration techniques to create MOSFETs on Si wafers that are composed of different semiconductor channels. Figure 5.3 shows two important examples, namely, a III–IV *n*-channel MOSFET based on InGaAs and a *p*-channel MOSFET based on Ge that can be integrated on the same wafer to take advantage of the increased electron and hole mobilities, respectively, for CMOS circuits. Further improvements in material quality and processing of these alternative semiconductor channel MOSFETs could likely lead to their commercial adoption within 5–10 years.

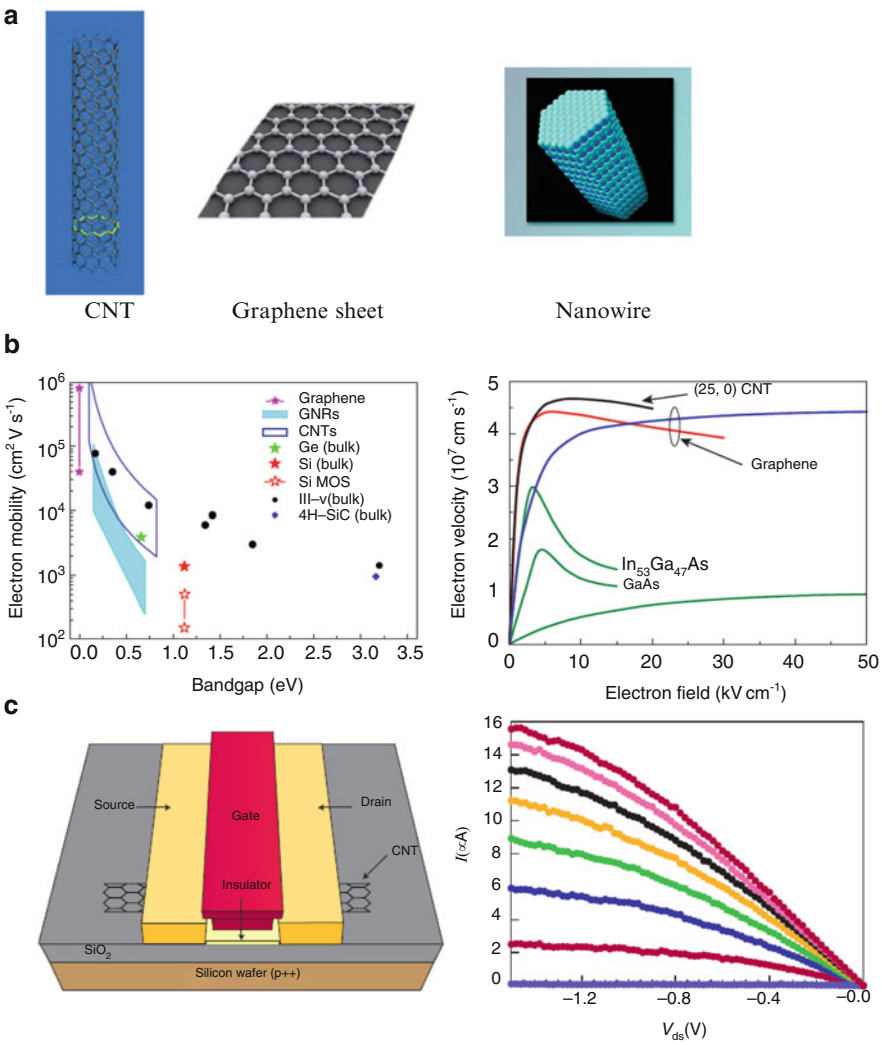


Fig. 5.4 Nanostructure-based MOSFETs. (**a**) Illustrations of a carbon nanotube, graphene, and semiconductor nanowire (image from Lawrence Berkeley National Labs). (**b**) CNT and graphene electron mobility data. (After F. Schwierz, Nature Nanotech. **5**, 487 (2010).) (**c**) CNTFET device schematic and drain current characteristics. (After [2])

Looking ahead even further, nanoscale MOSFETs based on nontraditional materials offer tremendous opportunities. In particular, three prominent candidates that are receiving a great deal of attention are carbon nanotubes (CNTs), graphene, and semiconductor nanowires. These nanostructures are illustrated in Fig. 5.4a. CNTs and graphene are 1D and 2D carbon-based nanostructures, respectively, which share many desirable electrical characteristics including very large carrier

mobilities and ballistic behavior along with excellent thermal and mechanical properties inherent in the C–C covalent bonding structure. Nanotubes have the additional capability of being either metallic or semiconducting (with a tunable bandgap) depending on the tube diameter and the way it is wrapped.[4] The 2D graphene sheet nominally behaves as a "zero-band gap" semiconductor; however, an effective band gap can also be created due to quantum confinement in thin strips of graphene known as graphene nano-ribbons (GNRs).[5] Figure 5.4b compares the electrical properties of CNTs and graphene to some standard semiconductors. Semiconductor nanowires, on the other hand, are one-dimensional structures that can in principle be based on either elemental or compound semiconductors as required.

Figure 5.4c shows a typical CNTFET device structure wherein the FET channel now consists of a semiconducting (~1–2 nm diameter) nanotube. Such devices can take various forms but usually display very high performance (drain currents, transconductance, etc.) compared to MOSFETs based on conventional materials. Similar devices have been realized using nanowires and GNRs. Nanowires and nanotubes share the additional advantage that they are amenable to improved electrostatic control of the channel via so-called gate all-around processes whereby the gate wraps around the typically cylindrical 1D wire channel creating more efficient and robust MOSFET switches.

Although FETs based on alternative nanostructures are still at an early stage of development, their excellent performance demonstrates strong potential for pushing MOSFET scaling well below the 10 nm level. At present these materials face important challenges related chiefly to large-scale integration and the control of their properties.[6] The "bottom-up" growth and processing of nanostructures such as nanotubes cannot be used to assemble viable high-density circuits at present. On the other hand, conventional top-down approaches (e.g., lithographic patterning) run into difficulties at very small dimensions and ultrahigh densities. In addition, controlling nanostructure structural and electrical properties consistently and with the high yields necessary for very large integrated circuits is an outstanding problem. Nevertheless, CNTs and related low-dimensional structures are seen as important to future nanoelectronics due to their superior combination of physical properties that may also eventually lend themselves to the creation of alternative devices and architectures beyond traditional CMOS (see discussion below).

[4] Single-walled carbon nanotubes are labeled using the notation (n, m), which defines the tube diameter and wrapping angle starting from a flat sheet of graphite.

[5] Planar or 2D graphene may also be useful for MOSFETs in analog/small-signal applications (e.g., microwave) where the channel does not need to be switched off completely.

[6] This is a general problem in nanoscale science and technology: How does one precisely control the placement, orientation, and structure of very large numbers of nanostructures?

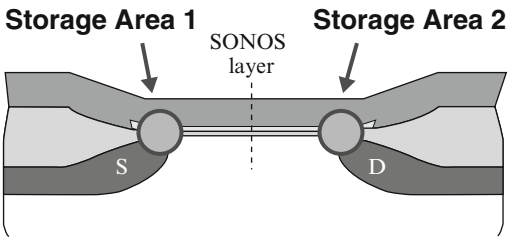


Fig. 5.5 Charge-trapping flash memory device structure. A so-called SONOS (silicon gate-oxide-nitride-oxide-silicon channel) structure is used to store charges on either side of a MOSFET channel as indicated by accelerating carriers toward the drain/source to write charge states into the nitride trapping layer, which modifies the threshold voltage of the device. Tunneling is used to erase the states (Adapted from Y. Polansky et al., IEEE International Solid-State Circuits Conference, 2006)

5.1.1.3 Memory Devices

The discussion above has focused mainly on MOSFET logic devices; however, memory must also continue to be scaled concurrently in order to achieve overall improvement in the performance of IC electronics.

In terms of DRAM, in order to continue to reduce the footprint of the 1T-1C memory cell, increased storage capacitance and decreased leakage currents are key factors. As with the gate dielectric for MOSFETs, the use of improved high-κ materials will continue with scaling of the DRAM storage capacitors in order to offset the reduced charge storage at smaller dimensions (due to less surface area). In addition, further improvements in high-performance access transistors such as multi-gate MOSFETs will be needed to reduce leakage and ensure that memory retention times remain adequate in future high-density DRAM circuits.

On the other hand, nonvolatile memory, predominately based on floating-gate flash memory devices (see Chap. 4), faces important challenges at very small dimensions that include insufficient floating-gate charge storage, memory retention problems for thin tunnel oxides, and cross talk between neighboring floating-gate cells. The introduction of high-κ material for the dielectric above the floating gate would allow adequate coupling during write/erase operations at lower voltage and help reduce cross talk. An alternative nonvolatile memory structure known as *charge-trapping flash* is another potential future option to mitigate cross talk while maintaining performance. In a charge-trapping MOSFET device (Fig. 5.5) only a single gate is used and the charge is stored in discrete traps within the dielectric layer instead of a floating gate. The lack of an intermediate floating gate in a charge-trapping device makes the gate–channel coupling less problematic than in standard flash memory devices while cross talk is reduced due to the localized charge storage within the dielectric stack. However, charge-trapping devices, by

themselves, do not address the fundamental problem of inadequate charge storage that will become a limiting factor for very high-density (~10–20 nm feature sizes) flash memory devices.

5.1.1.4 3D Integration/ICs

A major and growing trend for electronics in the future is three-dimensional integration. This can be seen starting from the basic device structures up to the circuit and chip levels. In terms of packing density, 3D integration increases the number of devices per unit area by employing the vertical dimension[7] and thus allows Moore's Law type scaling to continue even once lateral (or 2D) scaling has reached its limits. In addition, using the third spatial dimension can often lead to improvements and address challenges that might not otherwise be possible using standard planar integration. We have seen that there is already precedent for such ideas, for example, the transition to nonplanar DRAM capacitors in the 1980s and the more recent introduction of multi-gate or "3D" FETs. This is now becoming a more general aspect of integrated electronics.[8]

Increasing memory density is one area where aggressive vertical scaling is needed to overcome some of the problems associated with deep nanoscale dimensions outlined earlier. In particular, 3D flash memory cell approaches have been developed for the low-cost integration of multiple stacked memory layers to overcome the few-electron limitation inherent in very small memory cells. As illustrated in Fig. 5.6, 3D flash can consist of either charge-trapping or floating-gate structures. The relatively simple structure of charge-trapping devices along with their aforementioned advantages may become a decisive factor for future 3D flash memory integration. Similarly, access or pass transistor structures that are built vertically may help increase the density of DRAM cells in future technology.

Another example of 3D integration is the stacking of individual chips such as the through-silicon-via technique shown in Fig. 5.7a. A higher-level approach, chip stacking not only allows device integration density to increase but also allows finished chips with potentially very different functionalities to be integrated into a single vertical stack of thin semiconductor wafers.

If these trends continue to grow and expand, future IC scaling (and Moore's Law) may no longer be defined simply in terms of a node length (22 nm, 15 nm, etc.) but rather by the number of layers used in a 3D integrated circuit process.

[7] For example by stacking multiple device layers vertically.

[8] An analogy often used is that of a city that increases its density via the introduction of tall buildings (skyscrapers) with many levels as opposed to low-rise structures. Similarly, we may think of the flat electronic chip becoming more of a cube as 3D integration increases.

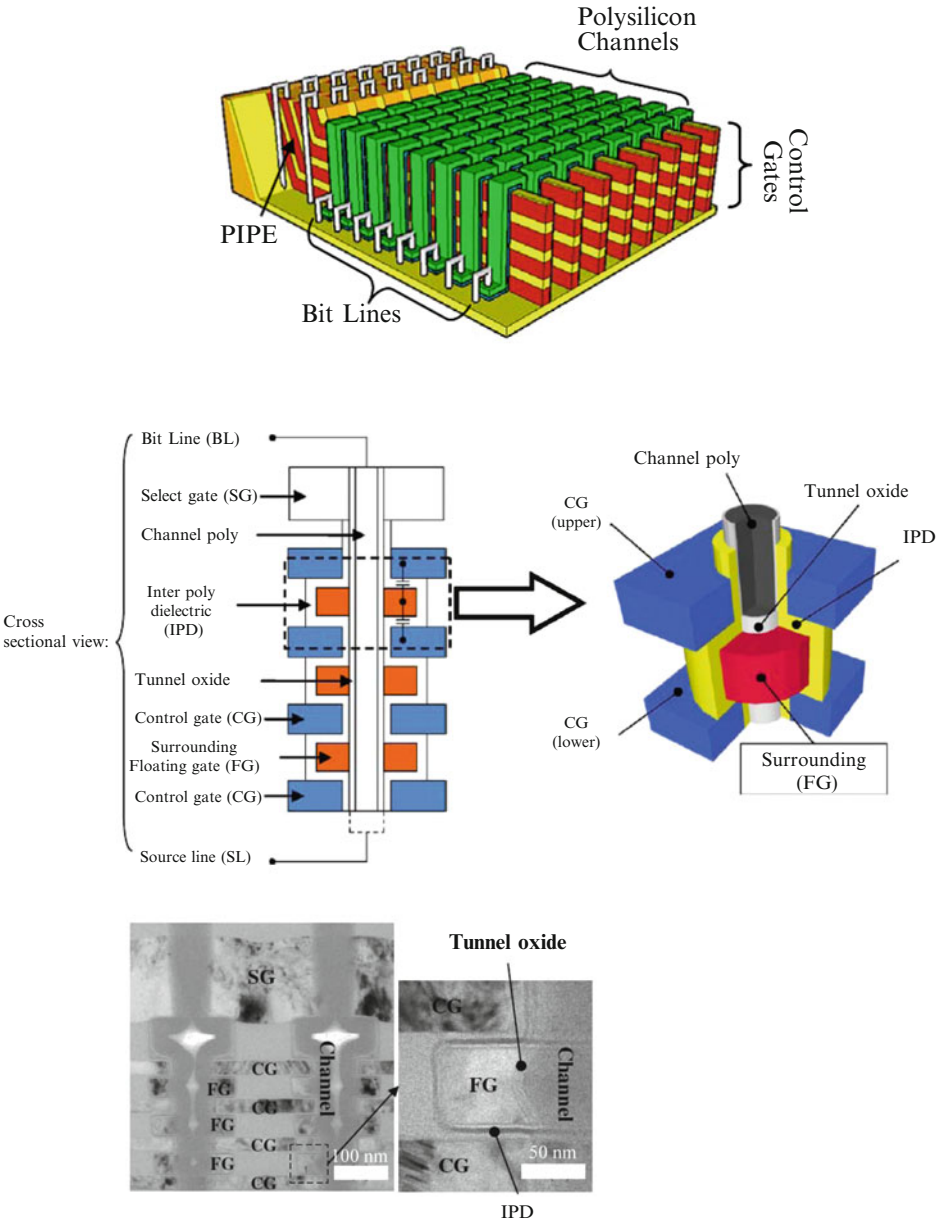


Fig. 5.6 3D flash memory structure examples. From *top* to *bottom*: Vertical channel stacked charge-trapping architecture. The active flash devices are based on the vertical channels (*green*) and control gate "pipes" (*red*). (J. Kim et al., Nanotechnology **22**, 254006 (2011)); 3D vertical floating-gate architecture based on cylindrical channel and surrounding floating gate with two vertically coupled control gates. (S. Aritome et al., Solid-State Electron. **79**, 166 (2013)). Vertical gate (horizontal channel) geometries have also been demonstrated

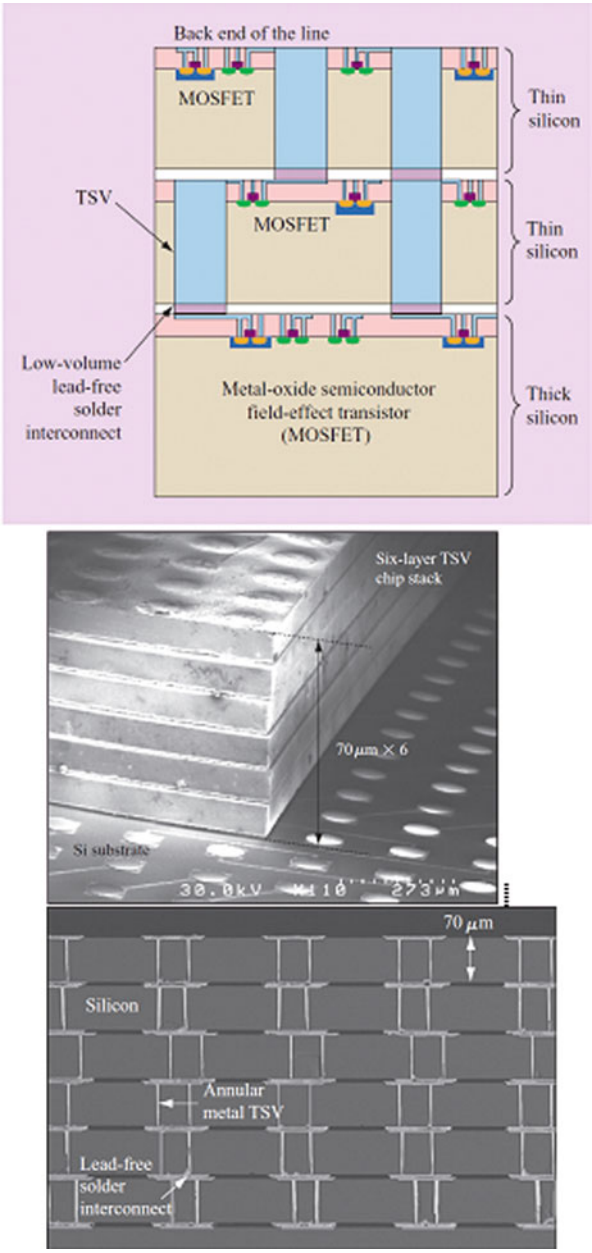


Fig. 5.7 3D chip-stacking process. In this example, through-silicon vias (TSVs) are used to interconnect multiple thin chips containing high-density MOSFET circuitry. Images of a six-layer stack are shown (After K. Sakuma et al., IBM J. Res. & Dev. **52**, 611(2008))

5.1.2 Alternative Devices and Architectures: Beyond CMOS

At some point, for dimensions below 10 nm, the standard MOSFET-based logic circuits and systems (i.e., CMOS) will likely be combined with and/or replaced by other devices that provide benefits and performance that either complement existing MOSFET technology or contribute new functionalities that might not otherwise be possible. The reason for this is mainly twofold: (1) MOSFET device scaling may not be optimal from a performance or cost point of view when channel lengths reach between 1 and 10 nm and (2) these dimensions, at the same time, provide fertile ground for other physical phenomena to provide devices based on new/different operating principles. These alternate approaches to information processing are thus usually referred to as *beyond CMOS* and can take many forms. This section briefly describes some of the main concepts for beyond CMOS devices and integrated circuit architectures.

5.1.2.1 Tunnel FETs

As discussed in Chap. 4, subthreshold conduction and in particular the subthreshold slope is an important parameter that impacts the performance of a MOSFET as a switching device. Instead of employing a standard type of junction between the source/drain and the channel, *tunnel FETs* (TFETs) on the other hand utilize gate-controlled band-to-band (or interband) tunneling between the heavily doped source and the channel[9] to turn the channel on and off (Fig. 5.8a). Both *n*-channel and *p*-channel TFETs can be made using this approach, thus allowing the realization of complementary logic circuits similar to CMOS. TFET devices are currently one of the top contenders for electronics that may initially appear beyond CMOS because they are very sensitive switching devices that can overcome the 60 mV/dec limit of standard MOSFETs,[10] which allows lower voltage operation compared to MOSFETs without sacrificing performance (Fig. 5.8b). An important challenge for implementing tunnel FETs is finding the appropriate material and device structure combinations that will provide sufficient output current for high-performance logic while maintaining the sharp subthreshold characteristics and associated power savings.

[9] This is similar to Zener tunneling (see Chap. 2) and the TFET structure also bears much in common with the original *tunnel* or *Esaki diode* (L. Esaki, Phys. Rev. **109**, 603 (1958)).

[10] Recall (Chap. 2) that this limit is due to the physics of charge transport (thermionic or diffusive) that can be encapsulated by the exponential (Boltzmann) factor in the ideal diode equation. Quantum mechanical tunneling on the other hand is governed by the width and height of the potential barrier with temperature typically playing less of a role (see Appendix A, Sect. A.1 for further details).

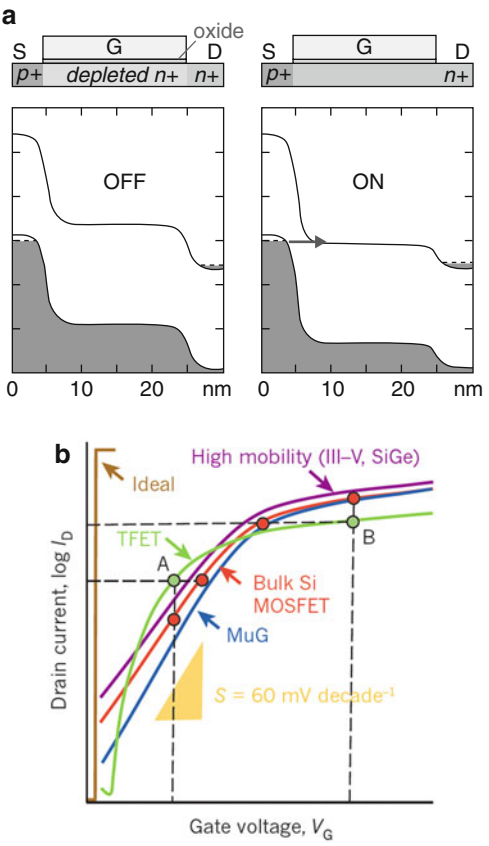


Fig. 5.8 Tunnel FETs. (**a**) Basic device structure and band edge diagram for an *n*-channel TFET below (*left*) and above (*right*) threshold. The analogous *p*-channel device consists of an n^+ source with p^+ channel and drain. (After [3].) (**b**) Comparison of subthreshold behavior of TFET to other types of FETs illustrating sharp turn-on characteristic (MuG = multi-gate FET) (After A. M. Ionescu, H. Riel, Nature **479**, 329 (2011))

5.1.2.2 Multistate Devices

Integrated electronics has been able to maintain most of its progress and success thus far by simply increasing the density of the same basic device function, i.e., the single bit or on–off switch, in order to perform increasingly complex logic functions in a finite amount of space (Moore's Law). The standard "spatial scaling" has been the norm for decades. A different approach to scaling instead involves changing the fundamental device function in order to achieve a higher density of information in a given amount of space. For example, a switch could have 3 or more states instead of the standard 2 states used for binary or Boolean logic. This type of "information scaling" through the use of multiple states or levels per device has already been realized commercially for nonvolatile flash memory in the form of *multilevel cell* devices that are able to store 2 or more bits of information (corresponding to 4 or more states) by using different amounts of charge on the floating gate (or charge-trapping layer). There is also precedent for multilevel

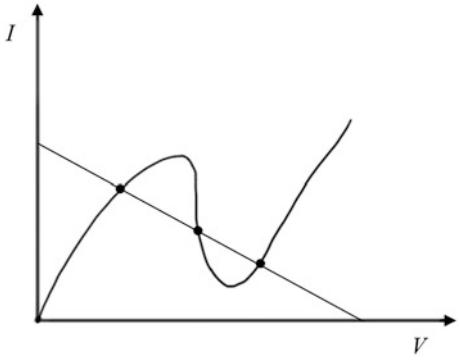


Fig. 5.9 Negative differential resistance. The load line of a simple circuit containing the negative differential device intersects the curve at multiple operating points as indicated, which can be used for switching between different states. Extending this concept to devices with multiple peaks allows multilevel logic to be realized

logic approaches based on research into *negative differential resistance* devices.[11] As illustrated in Fig. 5.9, the unique shape of the *I–V* curve in these devices allows for multiple stable operating points to be realized, which can be employed in multistate logic circuits. Such logic devices have not yet progressed beyond the development stage but serve to illustrate the general concept of increasing information capacity via multiple states.

Regardless of their specific implementation, devices that contain several states (whose number can be increased over time) may provide important future scaling opportunities for both memory and logic, which could lead to a type of "Virtual Moore's Law."

5.1.2.3 Alternatives to Electronic Charge Transport Devices

Modern solid-state electronics is based on the movement of charges in materials and devices via the application of electric fields, leading to currents and voltages that form the basic operation of electric circuits and systems. Electric charge can thus be considered the "state variable" of electronics that is used to represent signals and information. It is possible of course to consider devices based on other physical properties[12] as we already encountered in the early history of electronics and other areas of science/technology discussed briefly in Chap. 1. At the nanoscale, several

[11] The most common solid-state device implementation involves *resonant tunneling* structures. See e.g., S. Luryi, A. Zaslavsky in *Modern Semiconductor Device Physics*, S. M. Sze, Ed., Wiley, 1997, p. 253.

[12] For example, properties other than charge, or forces besides the electric field. Alternatively, interactions between several/many elementary particles can result in new collective properties.

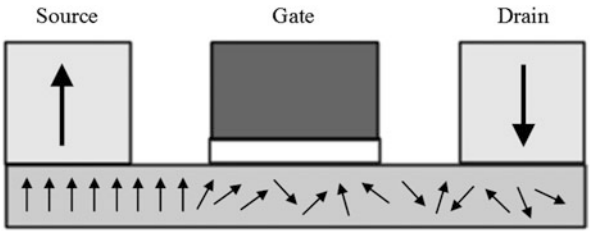


Fig. 5.10 Spin FET: Spin-polarized carriers entering the channel at the source are rapidly randomized by field-enhanced spin–orbit coupling under the gate, which leads to (spin-down) current at the drain (S. Datta, B. Das, Appl. Phys. Lett. **56**, 665 (1990))

alternative or unconventional state variables are being examined which may eventually extend or replace standard electric charge-based devices for certain applications or possibly provide completely new functionalities that were not previously possible using conventional electronics.

Spin

In addition to charge, electrons also possess a property known as *spin*, which can be thought of as an intrinsic angular momentum (see Appendix A). Electron spin can take one of two values, usually labeled "spin-up" and "spin-down." In conventional electronic devices the spin property of the electron is sufficiently randomized such that it does not play a direct role in determining device operation. However, if the spin state of an electron can be well defined, it could provide a different means of realizing useful device functions that are based upon spin instead of charge.

One well-known concept for a field-effect spin transistor is shown in Fig. 5.10. In this device structure the source and drain act as an effective source and sink, respectively, for polarized electron spins. Once the spin polarized carriers enter the channel the application of a gate voltage causes the spins to interact with the electric field and precess[13] and thereby lose their original spin orientation. In this manner the current through the device can be modulated by the gate voltage, similar to a conventional FET. One of the key advantages of such spin-based electronic devices is that they offer the potential for very low-power operation below the thermal voltage (k_BT/q)—something not possible with conventional electronics. Although appealing, spin electronics requires further progress in the efficient injection and extraction/detection of spins while maintaining adequate spin lifetime during transport in order to become viable.

[13] This is generally referred to as spin–orbit coupling.

Collective Excitations

Instead of considering single particle excitations such as the individual charges or spins of electrons, the interaction or combination of many particles can also lead to an array of interesting properties. Some well-known examples are ferromagnetic or ferroelectric materials, which possess intrinsic magnetic or electric fields (or polarizations), respectively. Collective excitations that consist of groups of particles acting in unison, often in the form of waves, are another case. Familiar examples include the electromagnetic field (photons) or lattice vibrations (phonons) but one can also speak of spin waves (magnons), charge-density waves, etc. Materials and devices based on these types of collective effects may lead to novel information processing devices and, from an electrical point of view, a group of n correlated particles can be treated as having a charge of nq and thus a lower corresponding thermal voltage constraint, namely, $k_B T/nq$.[14] A few approaches that utilize some form of collective interactions are discussed below.

The so-called *negative gate capacitance FET* utilizes a ferroelectric insulator material in the gate stack of a MOSFET (Fig. 5.11a) in order to improve the device electrostatics and thus allow lower voltage operation with a steep subthreshold slope compared to conventional transistors. The built-in polarization of the ferroelectric material increases the effect of the gate voltage in order to achieve a very sharp turn-on characteristic. Challenges that must be addressed include finding an effective ferroelectric material with repeatable switching behavior over many cycles and methods to reliably integrate these materials with standard silicon device processing.

In the Mott FET (Fig. 5.11b) the channel is composed of a strongly correlated electron material known as a Mott insulator that undergoes a gate-induced metal–insulator transition. The mechanism of switching is therefore an electronic phase transition. The primary advantages of a transistor based on such a phase transition include rapid switching and low-power operation with virtually zero leakage current, in principle. In addition, high levels of output current may be possible via this approach since the phase transition causes a large number of previously localized carriers to suddenly become mobile and participate in current flow. Further theoretical and experimental studies are needed before the full potential of the Mott FET can be realized.

Some fundamentally different models of information processing that do not rely on any type of FET structure per se can be found based on ferromagnetic materials: Fig. 5.11c illustrates the concept of *nanomagnetic logic* based on the interacting fields of neighboring small magnetic islands or dots. Here, the magnetization direction of each island is used to represent information and the interaction of different patterns of islands leads to logic functionalities. Note that in this case signals are transmitted solely by the magnetic field interactions and no currents are

[14] Other types of interactions in these "strongly correlated" materials can also lead to smaller potential energies than might otherwise be possible.

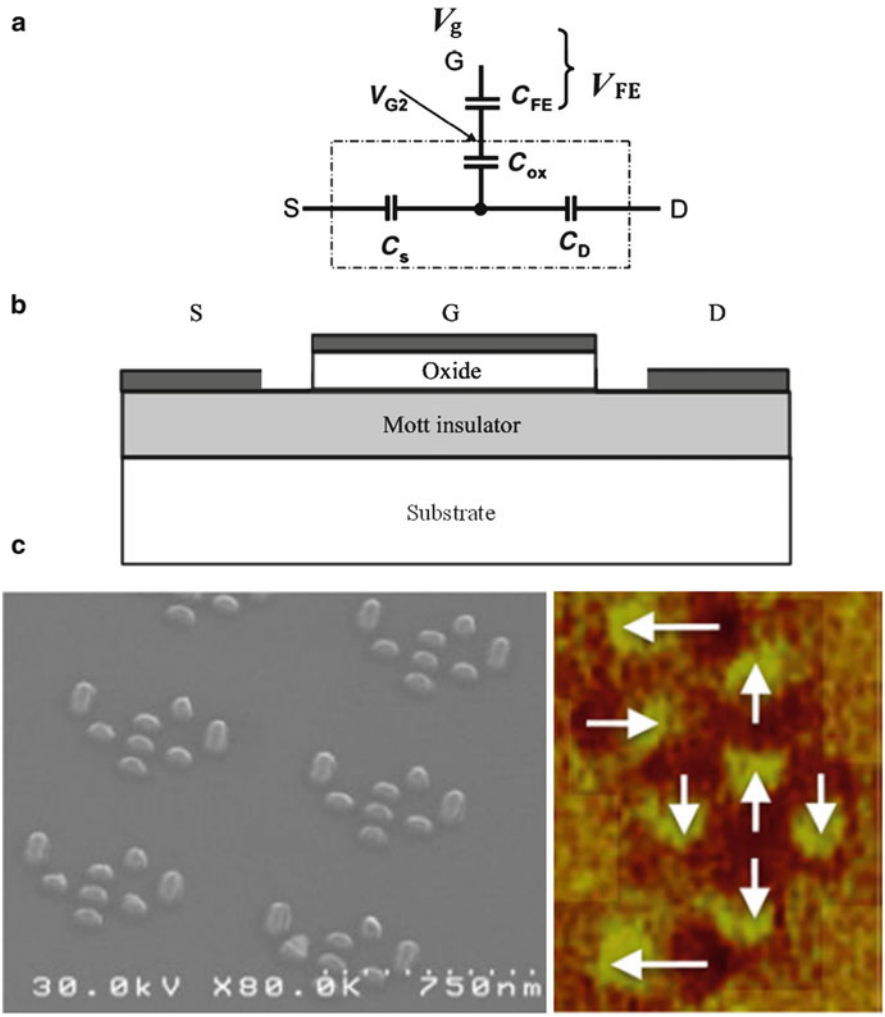


Fig. 5.11 Devices based on collective effects. (**a**) Equivalent circuit of a negative gate capacitance FET. (After S. Salahuddin, S. Datta, International Electron Devices Meeting, IEDM, 2008.) The standard MOSFET structure (*dotted box*) is modified by the addition of a negative capacitance ferroelectric insulator to the gate oxide stack that effectively amplifies the applied gate voltage. (**b**) Structure of a MOTT FET based on field-induced metal–insulator transition of the channel. (**c**) Nanomagnetic logic based on the in-plane magnetization of interacting magnetic islands. Different patterns can be used to perform logic functions based on the magnetization state of the islands. (After M. T. Niemier et al. J. Phys: Condens. Matter 23, 493202 (2011).) (**d**) Spin-wave logic concept. Electromagnetic induction is used to excite and detect spin waves that travel along a ferromagnetic spin waveguide. Multiple spin wave inputs can be combined to perform logic based on the phase of each input, for example (After A. Khitun, K. L. Wang, Superlattices Microstruct. **38**, 184 (2005))

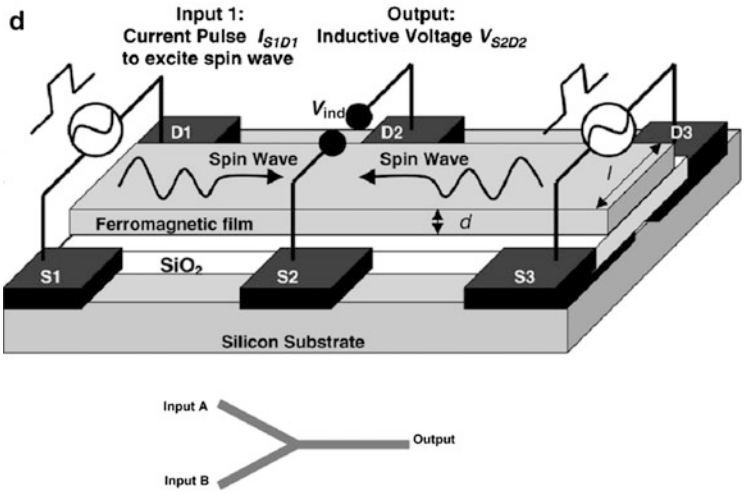


Fig. 5.11 (continued)

required, implying energy dissipation could be very small. *Spin-wave logic*, on the other hand, is based on collective spin excitations (magnons) that are used for transmitting and processing information. The two main components of spin-wave devices (Fig. 5.11d) are the input magnetic fields that excite the waves and the waveguides (spin-wave buses) that carry the waves from one point to another and also allow logic to be performed via constructive/destructive interference of multiple input waves. There are a number of unique aspects to such "wave-based computing" that are very desirable, including building logic devices with fewer active elements than conventional approaches and increased data/processing throughput by utilizing multiple frequencies or signals in parallel in the same device structure (multiplexing). Both nanomagnetic and spin-wave logic are still in the early development stages, but these and other novel concepts continue to be actively explored.

Electrochemical Switches

An electrochemical or *atomic switch* is essentially a conducting path that is created or destroyed by the movement of atoms during an electrochemical reaction. The switch formation process is illustrated schematically in Fig. 5.12a: Typically, metal ions form a bridge or fine filament between two electrodes to create the "on" state of the atomic switch and the reverse process breaks the connection and turning the switch "off." The switch can take both two- and three-terminal configurations (Fig. 5.12b) and experimental data on these devices appear very promising. In addition to large on/off ratios, these switches are also intrinsically nonvolatile and thus offer the possibility of logic circuits and systems that do not lose their state

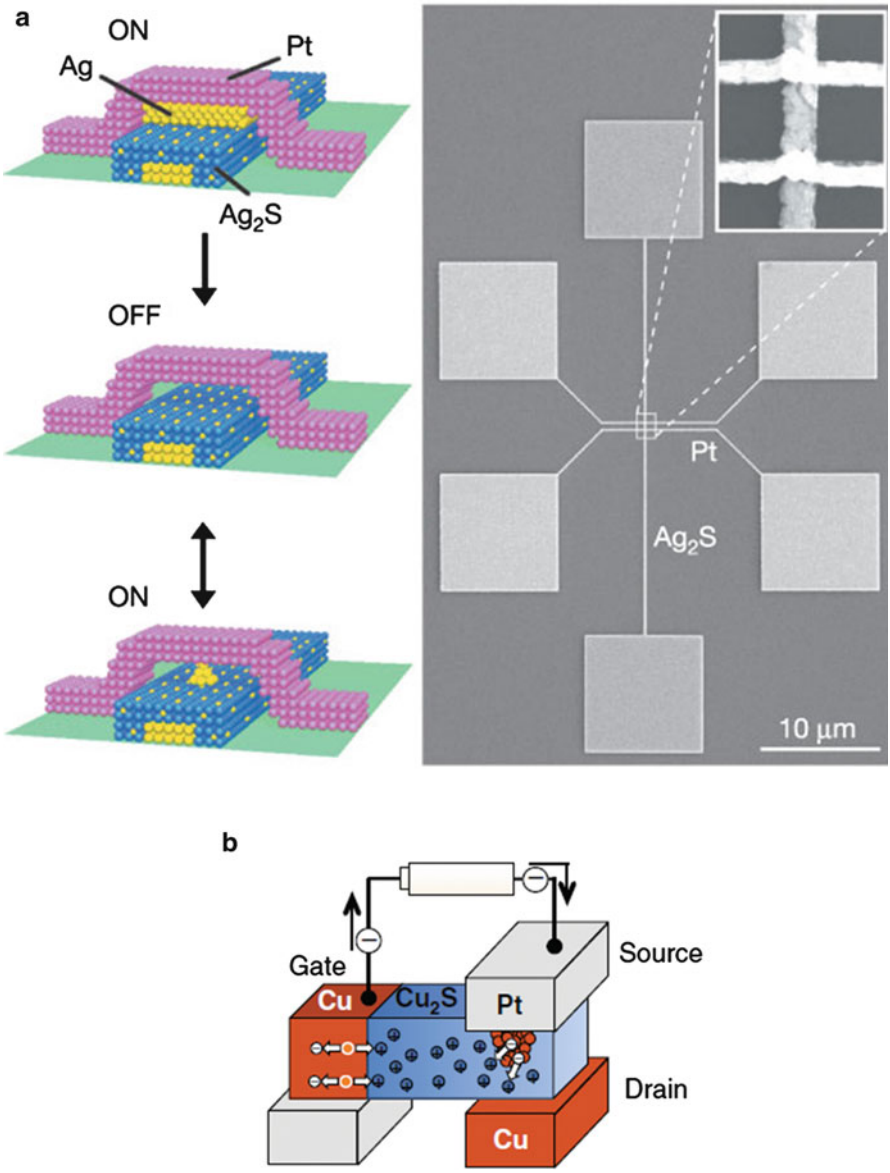


Fig. 5.12 Atomic switch. (**a**) Operating principle based on the formation/dissolution of a conducting atomic bridge formed via the application of a positive or negative bias to the top wire. Experimental realization of a cross-bar structure is shown in the microscope image. (After K. Terabe, T. Hasegawa, T. Nakayama, M. Aono, Nature **433**, 47 (2005).) (**b**) Three-terminal atomic switch in which a gate is used to control the movement of atoms in the channel (After T. Hasegawa, K. Terabe, T. Tsuruoka, M. Aono, Adv. Mater. **24**, 252 (2012))

when the supply voltage is turned off (nonvolatile memory using similar structures is described in the next section). Issues to be addressed or improved include switching speed, device uniformity/endurance, and a more detailed understanding of the basic device physics underlying the electrochemical switching process. Atomic switches are interesting in that they rely on chemistry and the movement of atoms to form very small wires, which can be thought of as nanoscale versions of the early chemical devices from the beginning of nineteenth century.

5.1.2.4 Emerging Memory Devices

In addition to beyond CMOS logic devices, several types of alternative memory structures are also being developed to expand/extend information storage beyond DRAM and flash. We focus here on two of the more promising options for future IC memory.

Resistive RAM

Metal–insulator–metal layered structures can be used to form nonvolatile RAM in a manner that is similar to the atomic switches discussed above. However, in the case of resistive RAM[15] the insulating layer typically contains charged species (ions) that can act as dopants to form conducting pathways or filaments in response to an applied electric field (Fig. 5.13a). Depending on the polarity of the applied voltage, the electrochemical formation or dissolution of the conducting pathways can be achieved, resulting in a low- or high-resistance state that defines the two values of the memory cell. Since resistive RAM depends on a resistance change (and not a capacitive effect) it is believed that it can be scaled to very small sizes.[16] In addition, the relatively simple structure of the two-terminal memory cell implies that it can be made at low cost and scaled to very high densities. As with atomic switches, a clear understanding of the physical mechanisms governing switching in resistive memory is a key challenge for technological implementation.

Phase Change Memory

Phase change RAM is a nonvolatile memory that relies on the resistivity difference of two different *structural phases* of a material. In particular, the difference in the resistance of an amorphous state (large resistance) and crystalline state (small resistance) of materials known as chalcogenide glasses is used to store the two levels required for a binary memory bit. As shown in Fig. 5.13b the phase change

[15] Often referred to as *redox RAM* due to the electrochemical reaction involved.

[16] Smaller sizes should also allow faster operation (due to ion transport over smaller distances).

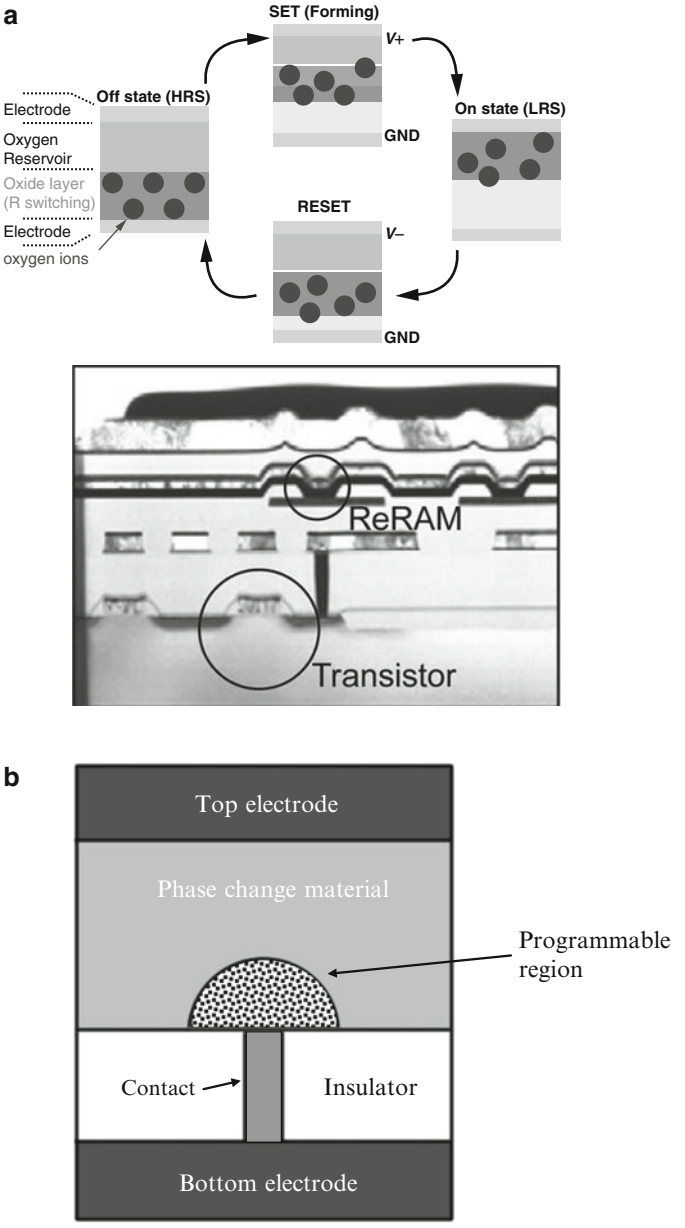


Fig. 5.13 Emerging memory devices. (**a**) Resistive RAM states based on metal–oxide–metal structure. Prototype IC cross section shows resistive RAM cell integrated with an access transistor. (After H. Akinaga, H. Shima, Proc. IEEE **98**, 2237 (2010).) (**b**) Phase change memory cell structure. The bottom electrode contact consists of a narrow high-resistance "heater" that can alter the crystalline structure and thus resistance in the region indicated in order to program the memory state of the phase change cell

memory cell consists of two electrodes with a phase change layer sandwiched in between. Voltage pulses are used to either melt (randomize, shorter pulse) or anneal (crystallize, longer pulse) the phase change layer. In order for phase change RAM to become the memory of choice for high-density integrated circuits the relatively large currents and long anneal (set) times must be addressed. However, since phase change RAM is a fairly well-established technology that is maturing (and being subjected to scaling) it is anticipated that its performance will improve upon reaching nanoscale dimensions and early experiments support this hypothesis.

5.1.2.5 Emerging Information Processing Architectures

An important aspect of nanoelectronics, and in particular beyond CMOS logic and memory, relates to the way these components are combined to create systems for processing information. Conventional digital (binary) logic based on CMOS technology utilizes what is known as a von Neumann architecture or similar schemes, wherein a set of instructions stored in memory (i.e., a program) is executed or "run" sequentially by the logic circuit(s) (or processor). In this way there is a clear separation between each logic operation or computation, which occurs in a serial manner. However, this may not be ideal (or at least not the only viable option) for many of the emerging device technologies discussed earlier.

For example, the nanomagnetic logic described earlier consists of a locally interconnected cellular automata architecture that supports parallelism. Another interesting aspect of many of the emerging beyond CMOS logic devices is that they can also be used for (nonvolatile) memory. This dual property implies that memory and logic could be combined into a single unit in future architectures. This could lead to, for example, nonvolatile computing applications, in addition to nanoscale programmable logic arrays and novel reconfigurable computing architectures (Fig. 5.14).

More generally, one can look beyond simply digital information processing to other models that may take advantage of the novel properties offered by alternative devices. For instance, collective excitations such as the spin waves described above could be employed to create a distributed nanoscale computational network governed by nonlocal excitations. Carrying these types of ideas further could lead to information processing systems that utilize analogue or continuous signals, e.g., biologically inspired computation.

Whatever forms the eventual beyond CMOS technologies end up taking, it is likely that they will augment rather than completely replace state-of-the-art CMOS integrated circuits. Thus a post-CMOS hybrid type of electronics, where new devices are gradually integrated (similar to how improvements to MOSFET technology have been made), appears to be the most likely path electronics will take in the years beyond 2020.

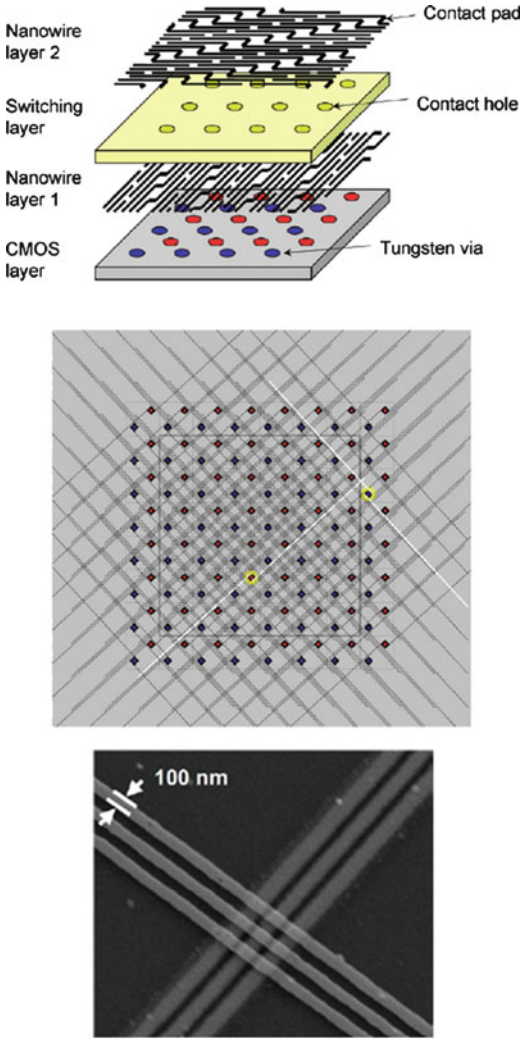


Fig. 5.14 Nanoscale crossbar programmable logic array combined with CMOS (After Q. Xia, et al., Nano Lett. **9**, 3649 (2009))

5.2 Microelectromechanical Systems

As mentioned at the beginning of this chapter, the influence of silicon electronics and the integrated circuit is extending beyond traditional electronics/information processing: The techniques and materials developed for ICs are being applied and adapted in new and diverse areas of application. Perhaps the most well-established extension of silicon planar processing thus far is the creation of micron-scale electromechanical devices on a chip, commonly referred to as microelectromechanical

systems or MEMS.[17] MEMS devices can take the form of various sensors, actuators, and/or energy-harvesting structures, ranging from very simple discrete devices to larger scale integrated and embedded systems with multiple interactions (electrical, mechanical, optical, chemical, etc.).[18] This relatively young field[19] has grown tremendously in the past couple of decades with devices that are now used in many products and industries including telecommunications, consumer electronics, automotive, medicine, military, aerospace, etc.

5.2.1 Materials and Structures

Figure 5.15a illustrates some of the fundamental structures used in MEMS design. These include cantilevers, bridges, fluid channels, rotating/torsional elements, membranes, etc. that can be integrated with various electrical components and circuitry. By borrowing, and sometimes modifying, techniques from integrated circuit fabrication, layer-by-layer process flows such as that shown in Fig. 5.15b allow a broad assortment of intricate electromechanical structures to be fabricated and integrated into compact systems. Silicon is the most common MEMS material, but many others including silicon nitride, polymers, glass, and metals are also widely used depending on the application.

The resonant frequency of a cantilever undergoing simple harmonic motion, perhaps the most basic MEMS structure, is given by the usual expression

$$\omega_0 = \sqrt{\frac{k}{m}} \qquad (5.1)[20]$$

where k is the spring constant and m the mass. This first-order description of the cantilever shows that for smaller systems (i.e., those with smaller mass) the characteristic frequency of motion can be quite large (MHz–GHz frequencies are quite common)—much larger than what is normally associated with mechanical structures. This is important for the application and integration of small electromechanical devices since they can operate at typical electronic frequencies.

The microfluidic channel is another basic building block for MEMS wherein either *pressure-driven* (based on a pressure gradient) or *electro-osmotic* (based on

[17] In addition, nanoscale MEMS structures are usually referred to as nanoelectromechanical systems or NEMS.

[18] This is an example of the "More than Moore" approach briefly introduced in Chap. 4.

[19] The first MEMS devices began appearing around the mid-1980s and have now become a multibillion dollar industry. This is another example of modern advances in manufacturing for electronic ICs leading to a rebirth of sorts for technologies such as electromechanical devices that were originally established much earlier.

[20] Recall that this arises from solving the differential equation for the motion of a particle in a parabolic potential (i.e., a system with a restoring force given by Hooke's Law, neglecting damping).

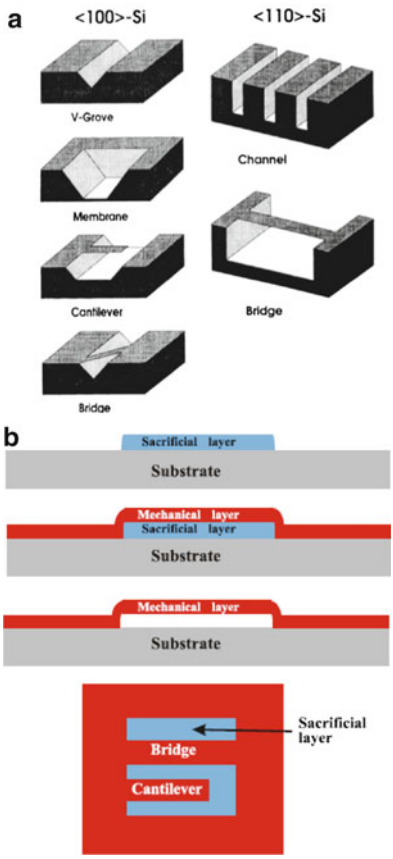


Fig. 5.15 (**a**) Basic MEMS structures; membranes/ beams, channels/ wells. (After W. Langi, Mater. Sci. Eng. **R17**, 1(1996).) (**b**) Example MEMS fabrication process flow (photolithographic patterning (see Appendix A) defines the lateral dimensions) (After P. J. French, P. M. Sarro, J. Micromech. Microeng. **8**, 45(1998))

an electrical potential gradient or applied electric field acting on mobile ions) flow is used to move a fluid through a microscale "pipe." In addition, liquid–surface interactions or capillary effects may also be important in small fluidic channels. Surface effects in general can be very important in MEMS because of the fact that the surface-to-volume ratio increases as materials are made smaller.[21] Therefore the effects of surface area in small structures can become very large.

5.2.2 Applications

5.2.2.1 Micromirror Arrays

One of the major applications for MEMS is the so-called digital micromirror display (DMD) used for image/video projectors and various other types of displays. The DMD module is a high-density array (~10^6 elements) of movable micromirrors

[21] For example, the surface-to-volume ratio of a cube with sides of length L scales as $1/L$.

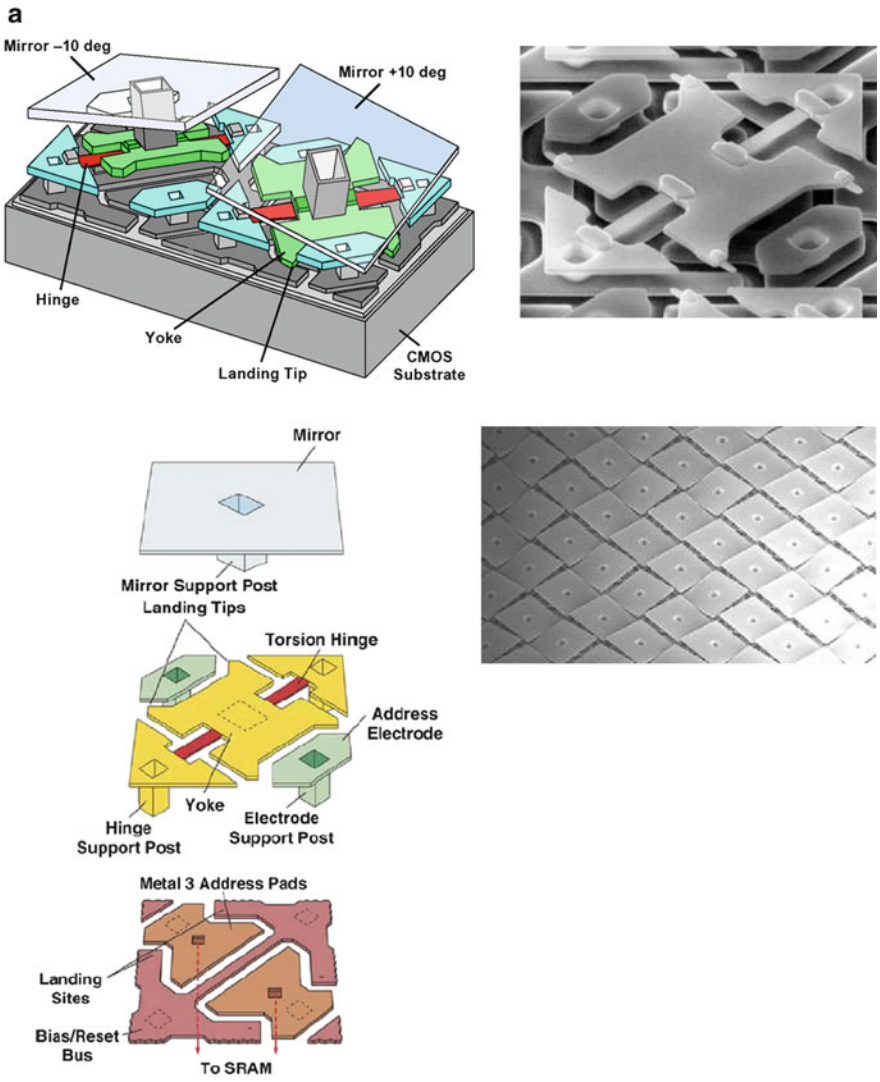


Fig. 5.16 Digital micromirror displays (Texas Instruments Inc.). (**a**) Individual mirror switching element structure and operation. (**b**) DLP chip and typical projector system configuration

that are integrated with CMOS circuitry and can be individually controlled. The mirrors are actuated by applying different voltages that changes the electrostatic attraction between the mirror and substrate as illustrated in Fig. 5.16a resulting in a torsional or twisting motion. Due to their small size (~10 μm × 10 μm) and the stiffness of the Si hinge/yoke on which the metal reflecting surface is placed, the mirrors are capable of switching thousand of times per second in order to create the desired output images. An example system based on the

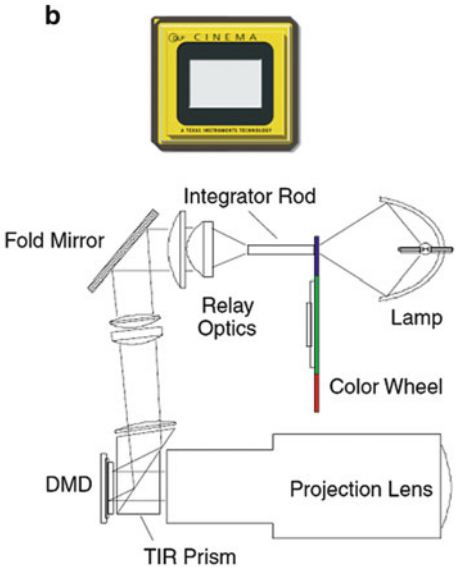


Fig. 5.16 (continued)

Texas Instruments DLP[22] chip is shown in Fig. 5.16b. Micromirror displays have proven to be robust and reliable display solutions that are in widespread use for consumer and large-scale digital projection applications.

5.2.2.2 Inkjet Printing

Another very widely used MEMS[23] device is the inkjet printer head, which employs microfluidic channels to deposit minute ink droplets on a printing surface in a controlled manner. The inkjet head typically consists of an array of small openings or nozzles in a silicon substrate connected to an ink reservoir. The two main methods used to eject the ink droplets are thermal printing or piezoelectric printing, as illustrated in Fig. 5.17. In both cases the ink nozzles are integrated with control electronics to create the inkjet chip. Advances in MEMS fabrications have enabled thousands of ink nozzles to be integrated in a small area on a semiconductor chip for fast, high-resolution printing with sub-pL droplets resulting in resolutions of up to 1,000s of dots per inch.

[22] Texas Instruments first introduced DMD products in the mid-1990s, under the digital light processing (DLP) brand.

[23] Note that MEMS usually encompasses both electromechanical and/or fluidic devices, whether or not they actually contain movable mechanical elements.

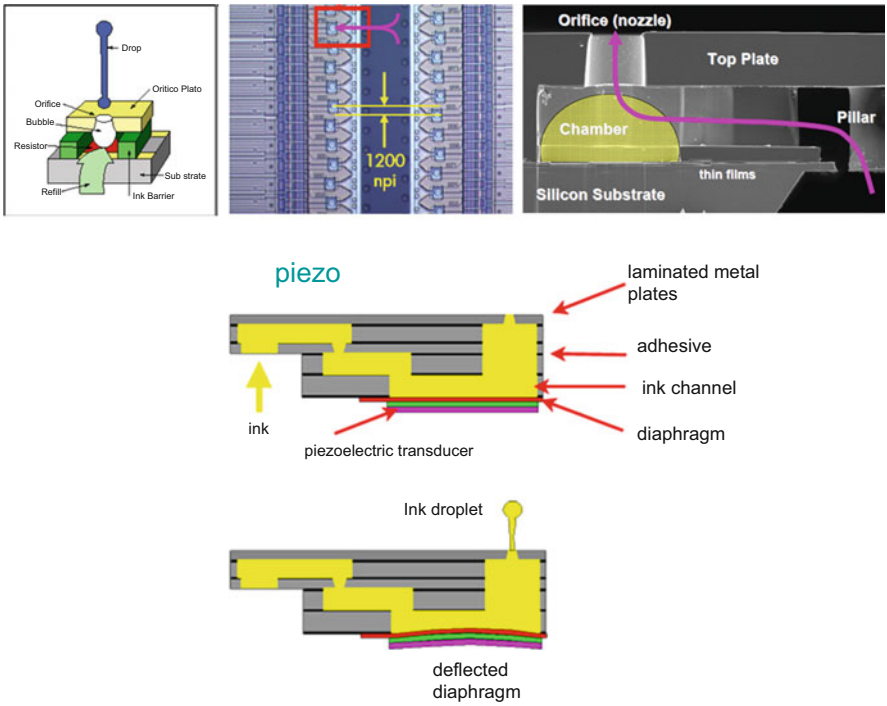


Fig. 5.17 Inkjet printing. In the thermal printing process (*top*) a thin film resistor is used to locally heat the ink until it expands to release a small drop through the nozzle as shown. Capillary forces refill the chamber by drawing in more ink from a reservoir after each drop is released. Piezoelectric printing (*bottom*) relies on the motion of a voltage-activated diaphragm to squeeze out the ink droplet (*Source*: Hewlett-Packard Company)

5.2.2.3 Drug Delivery

Microfluidic devices are also used for biomedical applications where the accurate controlled release of a substance in a miniaturized cost-effective package is very desirable. For example, microneedles are used for transdermal delivery of drugs with reduced skin/tissue damage and little or no pain compared to conventional needles. Typically an array of microscale channels (Fig. 5.18a) is integrated into a patch-like structure and used to deliver the therapeutic agent through the skin barrier. Similar types of MEMS devices for controlled drug delivery include inhaler nozzles (for more uniform aerosol size distribution) and micropumps (Fig. 5.18b).

5.2.2.4 Sensors

Sensing applications of MEMS form a large and diverse range of products and functions including accelerometers, gyroscopes, pressure sensors, and microphones, some of which are depicted in Fig. 5.19a. Many of these sensors feature

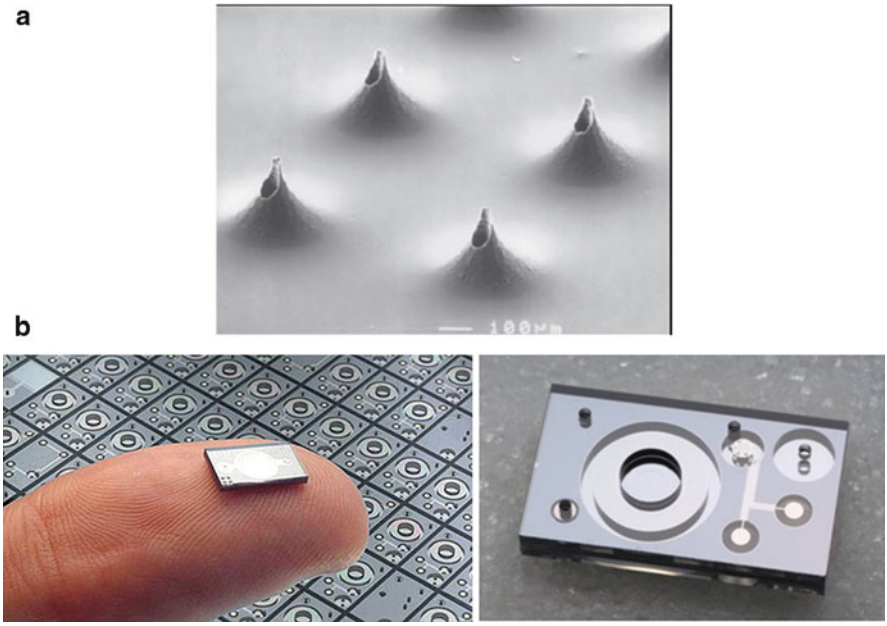


Fig. 5.18 Microfluidic drug delivery systems. (**a**) Electron microscope image of a microneedle array. (After B. Stoeber, D. Liepmann, J. Microelectromech. Syst. **14**, 472 (2005).) (**b**) Disposable microscale pump used for controlled delivery of insulin and other therapeutic agents (STMicroelectronics Inc.)

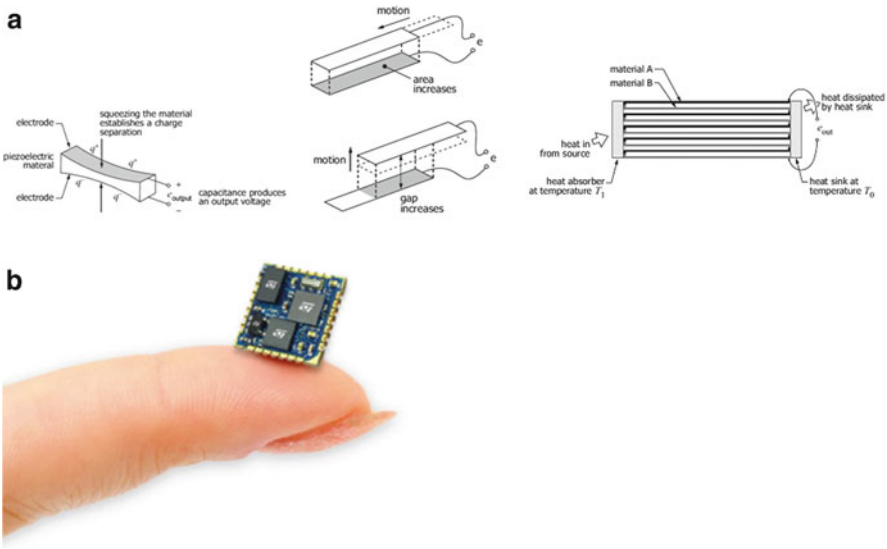


Fig. 5.19 (**a**) Examples of MEMS sensor operating principles. From *left* to *right*: piezoelectric pressure sensor, capacitive motion sensor for measuring acceleration, thermopile used for energy-scavenging of waste heat. (After [6].) (**b**) 9 DOF sensor integrated into a single package (STMicroelectronics Inc.)

in automobiles and consumer electronics, particularly mobile devices. An important trend is the integration of several MEMS sensors in order to reduce cost and increase performance. In particular, the development of inertial measurement units containing a 3-axis accelerometer, a 3-axis gyroscope, a 3-axis magnetometer[24] (or compass), and an ambient pressure sensor, so-called 10 degree of freedom sensors, is a major goal for the MEMS industry with a multitude of potential product applications. At present up to 9 degrees of freedom can be integrated at the package and/or chip levels (Fig. 5.19b).

5.3 Biochips

Integrated systems based on semiconductor technology are increasingly being used as diagnostic tools for the identification of biochemical substances and molecules in order to assist with the basic understanding and treatment of diseases. As with other applications, miniaturization allows an increase in performance (speed/function) while reducing cost. For point-of-care clinical tools integrated systems have the added benefit of requiring much smaller sample volumes. In this context, biological chips or *biochips* refer to integrated structures assembled on a semiconductor wafer or other solid substrates, which directly depend on biomolecular species for their device operation. This area is likely to grow tremendously over the next 10 years as the push toward "personalized medicine" based on an individual's genetic and biomolecular makeup becomes more ubiquitous.

5.3.1 Biomolecular Arrays

Many traditional biomedical diagnostics are based on the interaction or affinity of a molecular species (probe) with a target molecule contained in the biological sample (e.g., serum), which allows the amount or concentration of the target to be determined. The integration of this idea onto a chip containing an array of many probe molecules leads to the concept of a microarray or biomolecular array (Fig. 5.20a). Such bioarrays allow the sensing of many molecular species in parallel compared to more conventional approaches.

5.3.1.1 DNA Microarrays

The most well-developed and widely used bioarrays thus far are DNA microarrays for the study of genomes. DNA microarrays are based on the hybridization or bonding of complementary DNA strands. In a typical experiment, strands with known sequences

[24] See Hall effect discussion in Appendix A, Sect. A.2.

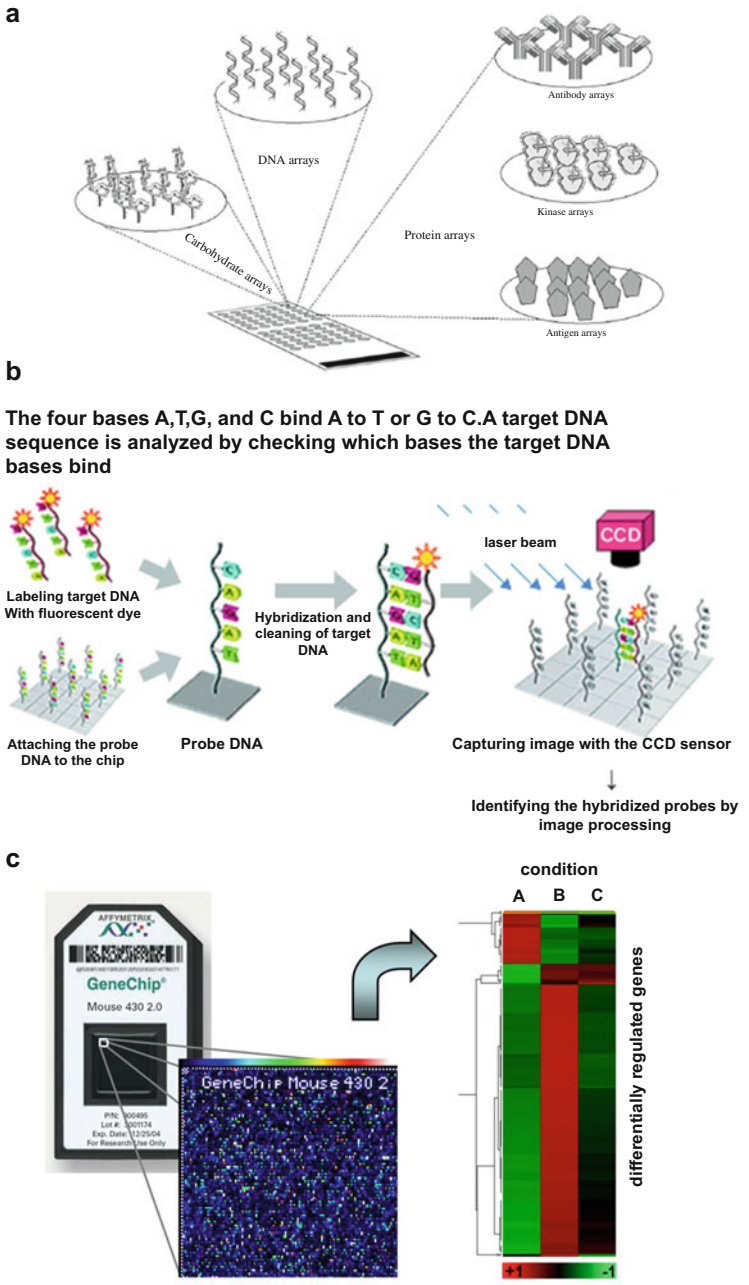


Fig. 5.20 Bioarrays. (**a**) General concept of a bimolecular array chip. (After A. Lagraulet, J. Lab. Automation **15**, 405 (2010).) (**b**) DNA chip principle. (**c**) Commercial gene chip image after hybridization with target strands (Affymetrix Corp.)

are first attached to a surface followed by exposure to fluorescently labeled DNA strands (the sample that contains the target DNA to be analyzed). If the target strands have the appropriate base-pair sequence they will attach or hybridize to the probe strands on the surface. After thorough rinsing and exposure to a light source only the locations with hybridized DNA should be visible (see Fig. 5.20b). By using a series of photolithography steps to fabricate the DNA microarrays, it is possible to achieve a very high density of unique single-stranded DNA[25] "spots" on a substrate (glass, silicon, etc.) and thus allow the analysis of up to millions of DNA sequences in parallel. An example of a commercial DNA or gene chip is shown in Fig. 5.20c. Such devices are now a standard tool for genomics.

5.3.1.2 Protein Arrays

The bioarray concept is not limited to DNA and can be extended in particular to include the very large class of biomolecules consisting of different types of proteins. Proteins are directly responsible for the behavior of cells and thus protein screening is very important practically in the understanding, diagnosis, and treatment of disease. Protein arrays can take many different forms (antibodies, enzymes, DNA-protein, tissues, etc.), but thus far high-density and reliable protein chips are still facing challenges for commercial implementation relating to the complexities of protein behavior and analysis. This continues to be very actively investigated area and with further advances in the field of proteomics protein microarrays will become more widespread and have the potential to make a great impact.

5.3.2 Lab on a Chip

The combination of fluidic components with other MEMS devices and various sensors and electronics can be used to miniaturize entire biochemical laboratory experiments on a single chip. This lab-on-a-chip (LOC) concept allows the delivery, mixture/separation, and reaction of biochemical/molecular species within a very small package to create perhaps the ultimate performance integrated biomedical diagnostic tools. Although the focus here will be on biotechnology applications, the LOC approach has applications in many fields including chemical analysis, chemical engineering, drug discovery, environmental monitoring, food safety, etc.

[25] Since there are four types of nucleotides that form base-pairs in DNA, four lithography steps are needed for each level in the fabrication of the single-stranded DNA. This places a practical (cost, time) limit on the length of the DNA strands that can be created and analyzed via this approach (typically strands less than 50 nucleotides).

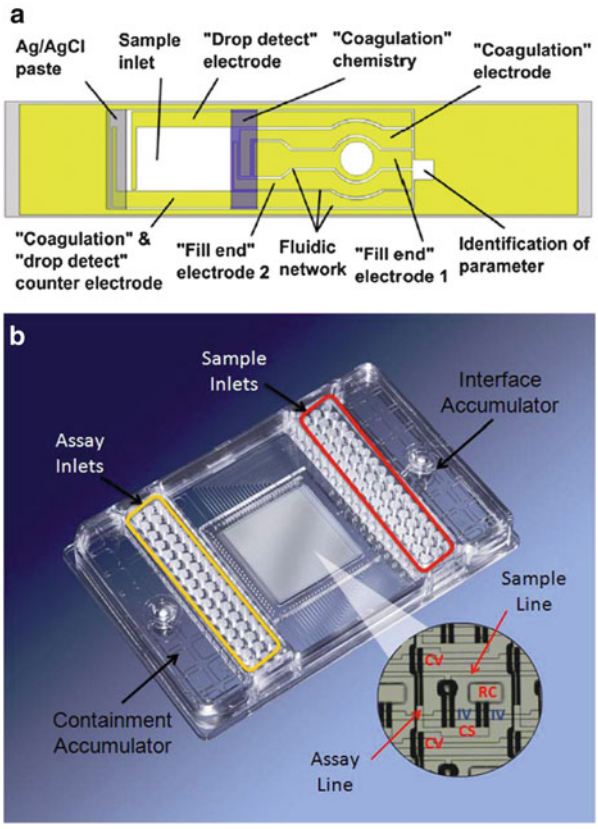


Fig. 5.21 Lab on-a chip designs. (**a**) Lateral flow test for blood coagulants. (After D. Mark, S. Haeberle, G. Roth, F. von Stetten, R. Zengerle, Chem. Soc. Rev. **39**, 1153 (2010).) (**b**) DNA microfluidic circuit (After J. Wang et al., BMC Genomics **10**, 561 (2009))

5.3.2.1 Immunoassays and Genetic Analysis

Many biological processes and conditions rely on the human immune response. Immunoassays are used to detect the presence of particular molecules or protein *biomarkers* present in biological fluids through their interaction with antibody molecules. These tests are commonly used in areas such as pregnancy, infectious diseases, heart attacks, and banned substance detection. LOC immunoassay devices can range from the fairly simple single-test variety (Fig. 5.21a) to more advanced devices that can detect a range of chemistries (e.g., glucose, sodium, potassium) in addition to biomarker molecules. The output of such devices is typically an optical signal or an electrochemically generated voltage.

An integrated microfluidic LOC platform for genetic analysis is shown in Fig. 5.21b. In these chips, a microfluidic matrix allows up to tens of thousands of DNA reactions to be performed in parallel resulting in significant time and cost savings. Another unique aspect of such integrated fluidic circuit platforms is that they are reusable or if necessary can be made disposable as well.

5.3.2.2 FET-Based LOCs

By using electronic components as the active biosensor elements in LOCs they could potentially achieve the same improvements in density and performance due to scaling that is found in electronic integrated circuits. In particular, modified MOSFET structures can be used as sensors to detect changes in their local chemical environment. The ion-sensitive field-effect transistor[26] (ISFET) is shown schematically in Fig. 5.22a. The operation of an ISFET is based on using a liquid electrolyte which contains the (charged) ion species to be detected as the *gate* of a MOSFET. Immersing the ISFET structure in an electrolyte will cause the potential on the liquid electrolyte gate to change depending on the concentration of ions present. Contact to the electrolyte gate is made using a reference electrode. By analogy with the MOSFET, recall from Chap. 4 that the threshold voltage is given by (assuming an *n*-channel device)

$$V_T = V_{FB} + V_C + 2|\phi_p| + \frac{1}{C_{ox}}\sqrt{2\varepsilon_s q N_a (2|\phi_p| + V_C - V_B)} \tag{5.2}$$

and that this determined the current flowing through the MOSFET based on the long-channel equations or their extensions (Sect. 4.2). The same equations are valid for the ISFET except that now the flat-band voltage, V_{FB}, is sensitive to the ion concentration in solution as the oxide surface effectively becomes charged[27] (c.f. Eq. 4.21). This leads to a change in the output current that can be detected or compensated for using the gate voltage in order to sustain a constant drain current. In either case a signal proportional to the ionic concentration is generated as the basic output of the ISFET sensor. A variety of ions can be detected using this approach including H^+, Na^+, K^+, Cl^-, etc., which makes these sensors useful for applications such as pH meters and the chemical analysis of biological samples. The size, speed, and cost advantages due to integrated circuit processing allow silicon ISFET chips to be used in multifunction point-of-care diagnostics applications as illustrated by the example in Fig. 5.22b.

An area where FET-based biochips have been recently applied is next generation DNA sequencing. Low-cost, ultrahigh-throughput sequencing technology is seen as the next evolution in genetics that will allow widespread access to "whole genome" analysis tools for medical purposes. The high density of sensors possible using ISFETs makes them a good candidate for this application. Figure 5.22c depicts the "ion torrent" sequencing technology: the basic sensing element used for sequencing is a modified ISFET structure consisting of a floating gate combined with a microwell in which multiple copies of the single-stranded DNA to be sequenced are deposited. The sequencing signal is generated using the "sequencing-by-

[26] P. Bergveld, IEEE Trans. Biom. Eng. **MBE-17**, 70 (1970).

[27] The particular dielectric layers used in an ISFET are chosen based on their affinity and sensitivity to the target ions.

synthesis" approach where the complementary base-pair sequence is formed sequentially, releasing an H^+ ion at each step along the strand when a base is attached. By monitoring the subsequent output signal of the ISFET the entire DNA strand can be sequenced. Each sensing element is combined into an array containing millions of elements along with the required electronic circuits for addressing and signal processing (similar to IC imaging or memory architectures)

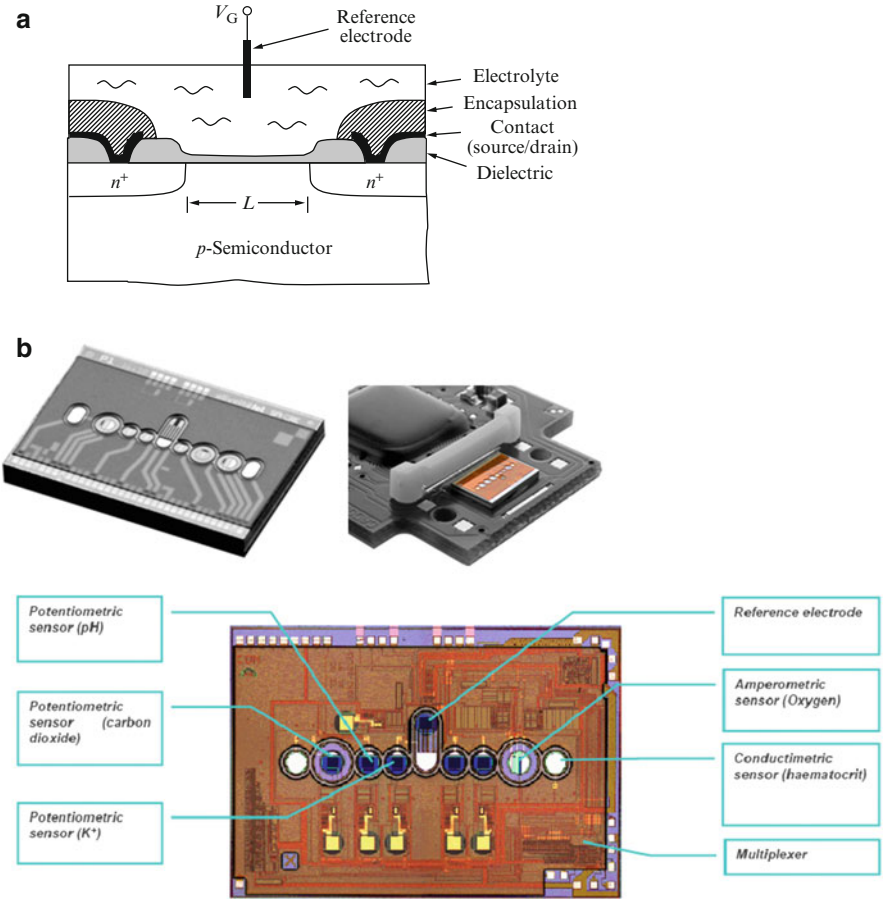


Fig. 5.22 FET-based biosensor chips. (**a**) Basic ion-sensitive FET device schematic. (After S. M. Sze, K. K. Ng, *Physics of Semiconductor Devices*, 3rd Edition, Wiley-Interscience, 2007.) (**b**) Commercial ISFET biosensor chip for blood analysis. (Sphere Medical Ltd.) (**c**) Ion torrent gene sequencing technology. (Life Technologies Corp.) An array of microwell-floating-gate ISFETs is integrated with a CMOS IC and exposed to microbeads containing the DNA strands to be sequenced. The sensor array is read out in a manner similar to IC memory. (**d**) Ion torrent sequence of Gordon Moore's genome (After J. M. Rothberg et al., Nature **475**, 348 (2011))

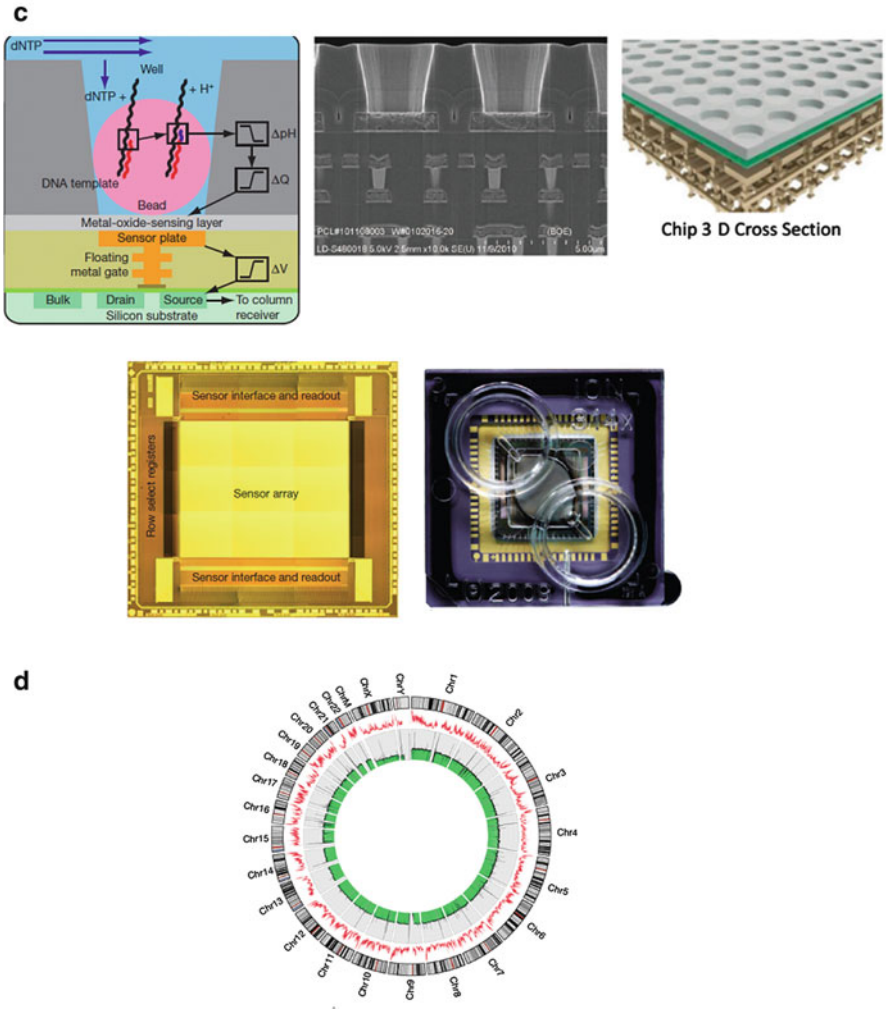


Fig. 5.22 (continued)

in order to allow high-throughput sequencing of many different strands in parallel. The all-electronic nature of ion-based FET sequencing is an important advantage that simplifies operation and lowers overall costs, in addition to allowing higher densities compared to other approaches that require optical readout. With continued scaling similar to Moore's Law in CMOS ICs greater parallelism with ion sequencing arrays approaching one billion elements or more[28] should also be possible.

[28] The output signal may become too weak at very small dimensions. This is similar to some of the issues discussed earlier for scaling memory to very high densities.

One of the initial genomes sequenced using the ion torrent approach was that of Gordon Moore himself (Fig. 5.22d).

As LOC and other biochip technologies develop further, the integration of emerging nanoscale devices and materials along with other types of sensors (optical, mechanical, thermal, etc.) will likely be realized in the future. These multifunctional systems will not only bring with them improved performance but also new possibilities and realms of application.

5.4 Conclusion

The technology behind the integrated circuit has led electronics down a tremendous path of innovation and success. That same technology is now branching out into many different areas bringing with it the lessons learned over several decades. This expansion of electronics and the integrated circuit will surely continue for quite some time and it should be clear from this chapter that the future of electronics is certainly very bright.

Lastly, it is perhaps worth noting that electronics indeed existed before the advent of the silicon integrated circuit and Moore's Law. It also seems certain that, in one form or another, whether the integrated circuit continues to evolve or reaches its peak, electronics will continue to play a major role in society for the foreseeable future. As for the exact form of electronic devices 50 years from now, I'll leave that to the experts:

> "I'm through with making predictions. Get it right once and quit."
>
> G. E. Moore

References

1. International Technology Roadmap for Semiconductors, ITRS, 2011–12; www.itrs.net
2. Avouris, Ph., Chen, Z., Perebeinos, V.: Carbon-based electronics. Nat. Nanotechnol. **2**, 605 (2007)
3. Seabaugh, A.C., Zhang, Q.: Low-voltage tunnel transistors for beyond CMOS logic. Proc. IEEE **98**, 2095 (2010)
4. Solomon, P.M: Device Proposals Beyond Silicon CMOS. IBM Research Report **RC24962** (2010)
5. Bernstein, K., Cavin, R.C., Porod, W., Seabaugh, A., Welser, J.: Device and architecture outlook for beyond CMOS switches. Proc. IEEE **98**, 2169 (2010)
6. Adams, T.M., Layton, R.A.: Introductory MEMS: Fabrication and Applications. Springer, New York (2010)
7. Bryzek, J., Roundy, S., Bircumshaw, B., Chung, C., Castellino, K., Stetter, J.R., Vestel, M.: Marvelous MEMS. IEEE Circuits Dev. Mag. **22**, 8 (2006)

Appendix A: Physics Primer for Electronic Devices

A.1 Quantum Mechanics

A.1.1 Particles and Wavefunctions: Elementary Wave Mechanics

We normally think of particles in a classical or Newtonian mechanics sense; that is, they occupy a specific point in space with well-defined position given by spatial coordinates and their velocity or momentum at any given instant is also precisely specified. In many cases, this picture is clearly satisfactory and very accurate. However, it is found that particularly when describing microscopic systems such as atoms and electrons it is more accurate to represent each particle by a *wavefunction* Ψ which describes its *distribution* or amplitude as a function of space[1] and time coordinates. For example, the plane wave is a harmonic function that is infinite in extent and periodic in both space and time

$$\Psi(x,t) = e^{i(kx-\omega t)} \tag{A.1}$$[2]

where $k = \frac{2\pi}{\lambda}$ and $\omega = 2\pi\nu$

are the wave vector (or wave number) and angular frequency, respectively.[3]

In quantum mechanics the basic equation of motion that governs the wavefunction is the *Schrödinger equation*:

$$-\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2}\Psi(x,t) + V(x)\Psi(x,t) = i\hbar\frac{\partial}{\partial t}\Psi(x,t) \tag{A.2}$$

[1] Most of the discussion will only involve one spatial dimension (e.g., x) for simplicity, but all the results can be generalized to three dimensions in a straightforward manner.

[2] Recall $e^{i\theta} = \cos\theta + i\sin\theta$.

[3] λ and ν are the wavelength and frequency, respectively.

C. Papadopoulos, *Solid-State Electronic Devices: An Introduction*,
Undergraduate Lecture Notes in Physics, DOI 10.1007/978-1-4614-8836-1,

where $\hbar = \frac{h}{2\pi}$

h is Planck's constant, $V(x)$ is the potential energy of the particle, m is its mass, and $i^2 = -1$.

The Schrödinger equation is a *separable* partial differential equation and its solutions can be shown to take the form

$$\Psi(x,t) = \psi(x)\exp\left(-\frac{iEt}{\hbar}\right) \tag{A.3a}$$

where $\psi(x)$ solves the *energy eigenvalue* equation[4]:

$$\left(-\frac{\hbar^2}{2m}\frac{d^2}{dx^2} + V(x)\right)\psi(x) = E\psi(x) \tag{A.3b}$$

for a particle with energy E.

The Schrödinger equation is also linear, so further wavefunction solutions can be constructed from linear superpositions of the basic solutions:

$$\Psi(x,t) = \sum_n a_n \psi_n(x)\exp\left(-\frac{iE_n t}{\hbar}\right) \tag{A.4}$$

Equation (A.4) gives the most general solution where n is an integer labeling the different solutions and scalar coefficients a_n. It can be seen that once the energy eigenvalue equation has been solved the complete solution for the wavefunction can be directly obtained, which makes the energy eigenvalues and eigenfunctions particularly important.

A series of examples will help to illustrate the above concepts.

Example A.1: Free Particle For a particle that is completely free and not subject to any potentials $V(x) = 0$ and we therefore obtain the following equations:

$$-\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2}\Psi(x,t) = i\hbar\frac{\partial}{\partial t}\Psi(x,t),\ -\frac{\hbar^2}{2m}\frac{d^2\psi(x)}{dx^2} = E\psi(x)$$

Solutions to these equations are the plane waves that we saw earlier in Eq. (A.1):

$$\Psi(x,t) = e^{i(kx-\omega t)}$$

[4] $\psi(x)$ is known as the energy eigenfunction.

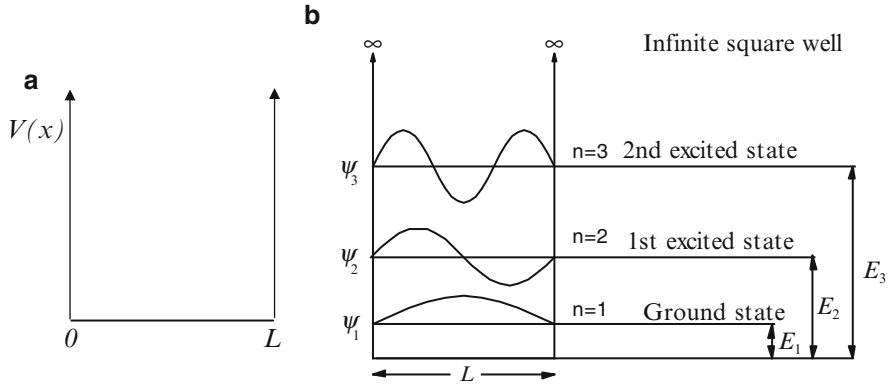


Fig. A.1 Infinite square well problem. (**a**) Potential well. (**b**) Energy eigenfunctions for three lowest energy states

where some important relations now arise as a result, namely,[5]

$$E = \frac{\hbar^2 k^2}{2m}$$
$$p = \hbar k = \frac{h}{\lambda}$$

where the latter expression is known as the *de Broglie relation*. In addition, it can be seen that

$$E = \hbar\omega = h\nu$$

(attributed to Planck and Einstein, respectively).

Example A.2: Bound Particle; Infinite Quantum Well In this case, the particle is completely confined to a small region of space. The bound particle can be modeled using an infinite square well potential shown in Fig. A.1a, often referred to as a "particle in a box" problem. Inside the well the potential energy is zero whereas at the boundaries it rapidly increases to infinity and the particle cannot escape or exist outside the well. Mathematically, for a well of length L, this corresponds to

$$V(x) = 0, 0 < x < L; \quad V(x) = \infty, \text{otherwise}$$

[5] Note that the relation $E = \hbar^2 k^2/2m$ does not apply to photons (i.e., massless particles). In general, we are also assuming that particles with mass are not traveling at very high velocities; in other words this is a nonrelativistic treatment of quantum mechanics and sufficient for most purposes when dealing with solid-state electronic devices. Also note that this equation is analogous to the classical expression for kinetic energy $E = p^2/2m$, where p is the momentum.

In order to solve the energy eigenvalue problem for this potential, we can note that inside the well the particle is not subjected to any potential and thus the solution should be the same as the free particle case we found earlier, i.e., a plane wave or in other words a combination of cosine and sine waves. However, in the case of the infinite quantum well we must also take heed of the boundary conditions at the edges of the well. Since the particle cannot exist outside the well we can write

$$\psi(0) = \psi(L) = 0$$

This means that the cosine function is not suitable (since it does not equal zero at the origin) and we can therefore write the energy eigenfunction as

$$\psi(x) = \sin kx$$

Now, applying the boundary condition at $x = L$ to this function yields the following relation:

$$\sin kL = 0 \Rightarrow kL = n\pi \Rightarrow k = \frac{n\pi}{L}$$

where n is a nonzero positive integer. Therefore by confining the particle to a "box" its energy levels become *quantized*:

$$E_n = \frac{\hbar^2 k^2}{2m} = \frac{\hbar^2 \pi^2 n^2}{2mL^2}$$

This is an important result. It shows that the energy levels are inversely proportionally to the width of the well squared; i.e., more confinement (smaller well) leads to higher energy levels. In addition, the energy levels become increasingly spaced out as n increases due to the factor of n^2. It also tells us the lowest energy a particle can have when confined to a region of space and that this *must* be a finite value, known as the ground state energy. Classically, we would expect a particle to eventually settle to the bottom of the well with zero energy. The quantum mechanical description of nature precludes this from happening and although the infinite square well is a simplified or "toy" model it serves to illustrate many concepts that occur in real systems and is often a good enough approximation to use for first-order calculations. In particular, the finite ground state energy is one of the reasons that matter is stable: without the wave behavior of particles, the electron and proton in a hydrogen atom, for example, would simply collapse into each other.

The energy eigenfunctions are explicitly (apart from a multiplicative constant, A to be determined),

$$\psi_n(x) = A\sin\frac{n\pi x}{L}, \quad n = 1, 2, 3, \ldots$$

and are plotted in Fig. A.1b, for the first 3 levels.

A.1.2 Interpretation of the Wave Function

The wavefunction itself is not something that can be directly measured. The standard interpretation of the wavefunction is based on the following:

$$|\Psi(x,t)|^2 dx = \begin{bmatrix} \text{probability of finding the particle} \\ \text{between } x \text{ and } x+dx \text{ at time } t \end{bmatrix} \tag{A.5a}$$

The magnitude of the wavefunction squared can be thought of as representing the "intensity" of the wave and represents the probability that a particle exists at a certain point in space. For particles which are *conserved*, in other words particles that cannot be destroyed, we require that the net probability for finding the particle somewhere in space must add up to 1, or

$$\int_{-\infty}^{\infty} |\Psi(x,t)|^2 dx = 1 \tag{A.5b}$$

known as the *normalization* condition for the wavefunction. For example, the wavefunctions of the particle in an infinite quantum well must obey

$$\begin{aligned} 1 &= \int_0^L |\Psi_n(x,t)|^2 dx = \int_0^L |\psi_n(x)|^2 dx \\ &= \int_0^L A^2 \sin^2 \frac{n\pi x}{L} dx \Rightarrow A = \sqrt{\frac{2}{L}} \end{aligned} \tag{A.6a}$$

Thus the complete wavefunction for the infinite square well problem is given by

$$\Psi_n(x,t) = \sqrt{\frac{2}{L}} \sin \frac{n\pi x}{L} \exp\left(\frac{-iE_n t}{\hbar}\right) \tag{A.6b}$$

Note that there is one important exception to the normalization condition: Plane waves cannot be normalized in the manner above because they have infinite extent. However, they are still very useful for calculations in which we are only interested in *relative* probabilities, as we will see later on.

A.1.3 Operators and Physical Quantities

In quantum mechanics all physically observable quantities are represented by operators. For example, we have already seen the energy operator in the energy eigenvalue equation:

$$\left(-\frac{\hbar^2}{2m}\frac{d^2}{dx^2} + V(x)\right)\psi(x) = E\psi(x) \Leftrightarrow H\psi(x) = E\psi(x) \tag{A.7}$$

The operator in brackets[6] representing energy is known as the Hamiltonian, H.

If the wavefunction Ψ of a particle corresponds to an eigenfunction of the operator representing an observable being measured, then that observable has a definite value. For example, the energy of a particle in the first excited state of the infinite quantum well has a precise value given by E_1.

A.1.4 Expectation Values

The expectation or average value of a quantity associated with a particle can be found from its wavefunction. For example, the average position of a particle is given by

$$\langle x(t)\rangle = \int x|\Psi(x,t)|^2 dx = \int \Psi^*(x,t)\ \ x\ \ \Psi(x,t)dx \tag{A.8}$$

which is essentially a weighted sum of the quantity of interest with the wavefunction probability distribution. Using this equation we can calculate the average position for a particle in the lowest energy level of the infinite quantum well as

$$\langle x(t)\rangle = \frac{2}{L}\int_0^L x\sin^2\frac{\pi x}{L}dx = \frac{L}{2} \tag{A.9}$$[7]

Note that in this case the average position is constant in time—this is a general property of energy eigenfunctions and they are referred to as *stationary states*. All expectation values of stationary states are constant in time.

In general the expectation value of some physical observable, Q, is given by

$$\langle Q\rangle = \int \Psi^*(x,t)\ \ Q\ \ \Psi(x,t)dx \tag{A.10}$$

A.1.5 Uncertainty Relations

A measure of the uncertainty or dispersion[8] of an observable associated with a particle is given by the *root-mean-square (RMS) deviation*:

$$\Delta A \equiv \left(\langle A^2\rangle - \langle A\rangle^2\right)^{1/2} \tag{A.11}$$

for an operator A.

---

[6] This can be thought of as the sum of kinetic plus potential energy just as in classical systems.

[7] Note that the expectation value of position for any energy eigenstate in the infinite quantum well will always be $L/2$ due to the mirror image symmetry about the center of the potential well.

[8] In this case dispersion refers to the spread of an observable about some average value.

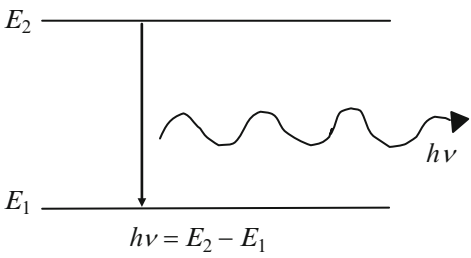


Fig. A.2 Excited state. Particle in upper energy level relaxes to ground state with emission of a photon

There are two very important *uncertainty relations* in quantum mechanics that arise from Eq. (A.11) and the properties of solutions to the Schrödinger equation:

The first relates the uncertainty between position and momentum, known as the *Heisenberg uncertainty principle*:

$$\Delta x \Delta p \geq \frac{1}{2}\,\hbar \tag{A.12a}$$

This oft-quoted relation tells us that we can never know the exact position and momentum of a particle at the same time.[9]

A different type of uncertainty that arises when dealing with quantum systems (such as lasers) is the *energy–time uncertainty relation*:

$$\Delta E\ \ \delta t \geq \frac{1}{2}\hbar \tag{A.12b}$$

Here, the "uncertainty" in time is not the same type as in the observables we looked at before, which arose from their probability distributions. (Time is a parameter in conventional treatments of quantum mechanics and does not have any inherent uncertainty.) Rather, δt is the time interval required for an *appreciable change* to occur in the properties of the system under study.

Example A.3: Energy–Time Uncertainty

Fixed stationary state We know that for an energy eigenstate $\Delta E = 0$ and thus the energy–time uncertainty relation yields $\delta t = \infty$[(i)]. In other words, the system does not change, as expected for a state fixed in time.

Excited state

Consider now the more interesting case of a particle that has been excited to some higher energy level and then decays. For a concrete example, suppose as shown in Fig. A.2 that a particle is in the second ($n = 2$) energy level of a quantum well and

[9] This should seem, along with many other results of quantum mechanics, somewhat counterintuitive. It may help to note that the scale set by Planck's constant is very small, which at least partly explains why everyday macroscopic objects do not normally display explicit wavelike behavior.

[(i)] Or vice-versa; $\delta t = \infty \Rightarrow \Delta E = 0$.

spontaneously decays to the lowest (ground) state by the emission of a photon. If the typical lifetime for this decay to occur is given by τ then the energy–time uncertainty imposes the condition

$$\Delta E_2 \geq \frac{\hbar}{2\tau}$$

After the particle reaches its ground state, no further decay can occur and we are back to a fixed stationary state which does not change in time, as in the previous case. In practice, the spread in energies determined by the energy–time uncertainty relation imposes a fundamental limit on how sharp the output spectrum of light-emitting devices can be.

Example A.4: Potential Barriers; Tunneling Consider a free particle with energy E impinging on a potential energy barrier[10] of width 2 L and height V_0, as illustrated in Fig. A.3a. By inspection, we can write the spatial part of the wavefunction in each of the three regions shown using by modifying the results for the free particle examined earlier, i.e.,

$$\psi_{\mathrm{I}}(x) = e^{ikx} + re^{-ikx}, \psi_{\mathrm{II}}(x) = ce^{-\kappa x} + de^{\kappa x}, \psi_{\mathrm{III}}(x) = te^{ikx}$$

where $k_{\mathrm{I,III}} = \sqrt{2mE}/\hbar$; $\kappa = \sqrt{2m(V_0 - E)}/\hbar$, $k_{\mathrm{II}} = i\kappa$ and it has been assumed that the incoming particle has unit amplitude.

In brief, these solutions are obtained by noting that in regions I and III the particle must have plane wave solutions since the potential energy is zero. In region I there is also a reflected wave in the opposite (negative) direction to the incoming particle, denoted by the coefficient r, whereas the plane wave traveling toward the right in region III has a transmitted amplitude coefficient t. In region II, known as the classically forbidden region, the kinetic energy is less than the potential energy and a classical particle could never exist in this region (it wouldn't have enough energy to overcome the potential barrier). However, quantum mechanically a plane wave solution can still be written as before, but now the wave vector becomes complex and the complex exponentials instead become regular (increasing and decaying) exponentials with coefficients c and d.

To solve for the coefficients, we match the wavefunctions in each region using the appropriate boundary conditions as required for solutions to the Schrödinger equation, namely, that the wavefunction and its first derivative must be continuous, i.e.,

$$\psi_{\mathrm{I}}|_{x=-L} = \psi_{\mathrm{II}}|_{x=-L} \qquad \psi_{\mathrm{II}}|_{x=L} = \psi_{\mathrm{III}}|_{x=L}$$
$$\frac{d\psi_{\mathrm{I}}}{dx}|_{x=-L} = \frac{d\psi_{\mathrm{II}}}{dx}|_{x=-L} \qquad \frac{d\psi_{\mathrm{II}}}{dx}|_{x=L} = \frac{d\psi_{\mathrm{III}}}{dx}|_{x=L}$$

which gives

[10] This can be considered as the quantum well problem flipped upside down.

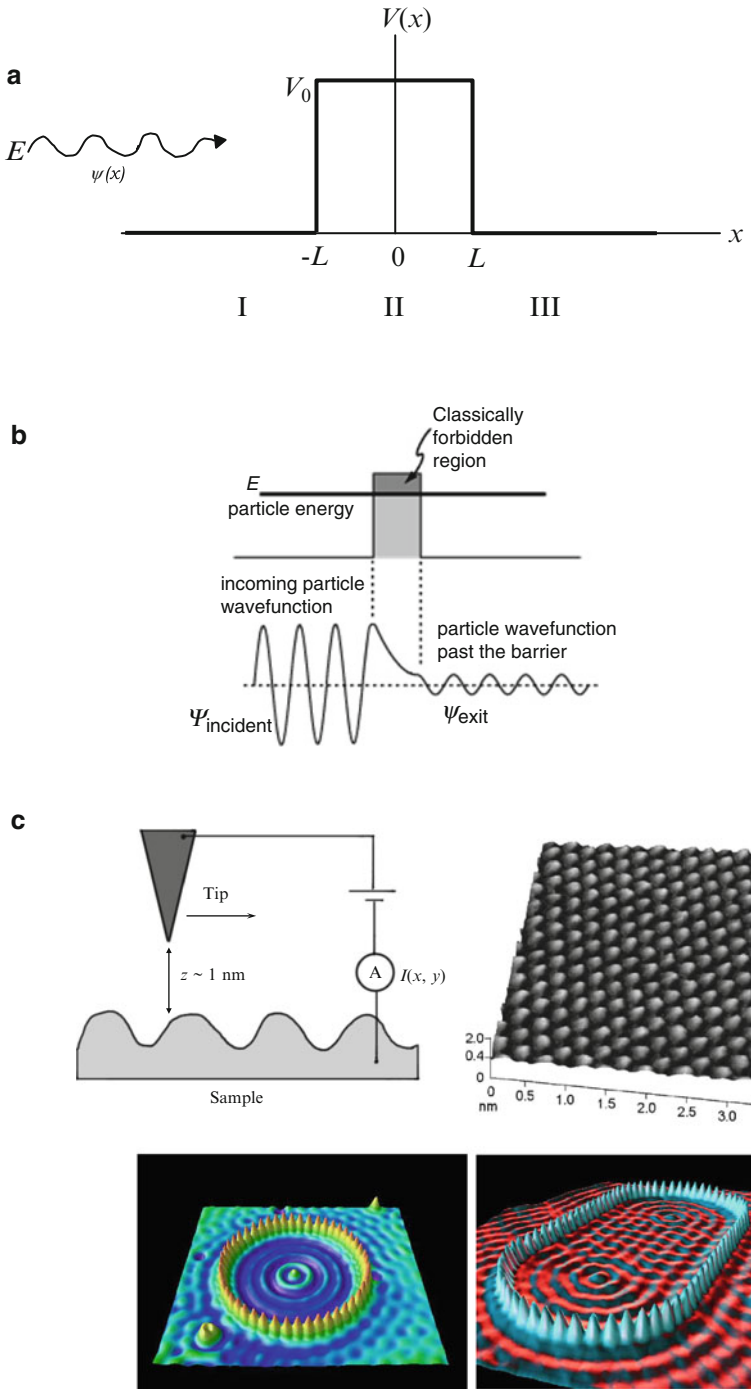


Fig. A.3 Potential barrier problem. (**a**) Square potential barrier with incident plane wave shown. (**b**) Resulting wavefunction at a given instant in time for a particle tunneling through the barrier. (**c**) Scanning tunneling microscope schematic and surface scan showing atomic resolution. Quantum corrals created by manipulating atoms on a surface to confine the electrons of a metal are also shown (IBM Corp.)

$$
\begin{aligned}
x = -L; \quad & e^{-ikL} + re^{ikL} = ce^{\kappa L} + de^{-\kappa L} \\
& ike^{-ikL} - ikre^{ikL} = -\kappa ce^{\kappa L} + \kappa de^{-\kappa L} \\
x = L; \quad & ce^{-\kappa L} + de^{\kappa L} = te^{ikL} \\
& -\kappa ce^{-\kappa L} + \kappa de^{\kappa L} = ikte^{ikL}
\end{aligned}
$$

These 4 equations can be solved for the 4 unknown coefficients. For example, the amplitude of the transmitted wave is

$$t = \frac{-4ik\kappa e^{-2\kappa L} e^{-2ikL}}{(\kappa - ik)^2 - (\kappa + ik)^2 e^{-4\kappa L}}$$

If κL is fairly large (in other words if the barrier is tall and wide) the probability of transmission,

$$T = \left|t\right|^2$$

can be approximated as

$$T \propto \exp(-4\kappa L)$$

Thus the particle has a finite probability of transmission through the barrier that depends sensitively on the height and width of the barrier. In particular, if the barrier becomes thin the transmission can be quite significant. This inherently quantum mechanical phenomenon is known as *tunneling*. The spatial part of the wavefunction for a particle tunneling through the barrier is sketched in Fig. A.3b.

A direct application of quantum mechanical tunneling is the scanning tunneling microscope (STM),[11] which employs the very sensitive dependence of tunneling with distance to form images with atomic resolution by scanning a sharp tip across a surface (Fig. A.3c). STM measures the tunneling current between a very sharp metallic tip (e.g., PtIr) and the surface being studied. The tip is scanned across the surface using a piezoelectric scanner(s) and its height is also controlled in a similar manner. Typically an order of magnitude change in current is observed for a 0.1 nm change in tip height and this results in the very high vertical resolution of STM ($<$1 pm). STM can also be used to manipulate individual atoms as illustrated by the well-known "quantum corrals" created by researchers at IBM.

[11] Invented by Binning and Rohrer at IBM in 1981.

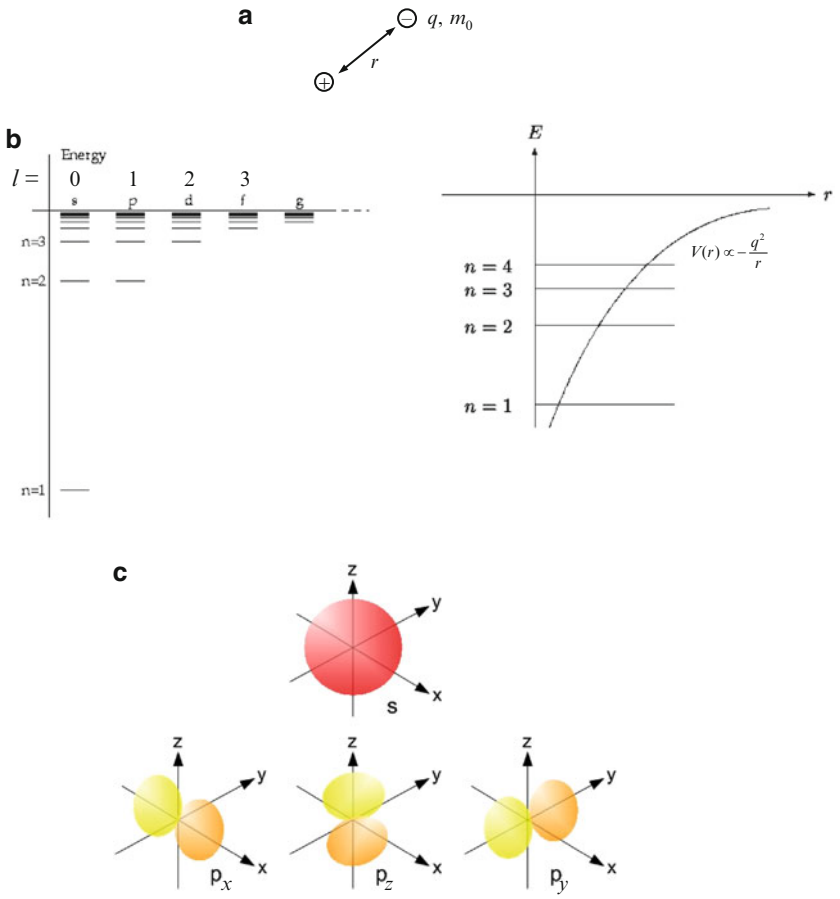


Fig. A.4 Hydrogen atom. (**a**) Negatively charged electron coupled to a fixed proton. (**b**) Energy levels versus n and l quantum numbers and Coulombic potential energy superimposed on energy levels. (**c**) Illustration of s- and p-orbital wavefunctions (different colors on p-orbitals correspond to positive and negative values of the wavefunction)

A.1.6 Atoms and the Periodic Table

Hydrogen Atom

The simplest atomic system, hydrogen, consists of one electron and one proton. The motion of the negatively charged electron about the positively charged fixed nucleus (Fig. A.4a) can be solved using the Schrödinger equation in 3D for the central (Coulombic) potential:

$$V(r) = \frac{-q^2}{4\pi\varepsilon_0 r} \tag{A.13}$$

where r is the distance from the origin (nucleus). The energy eigenfunctions and eigenvalues in spherical coordinates are given by

$$\psi_{nlm}(r,\theta,\phi) = R_{nl}(r)Y_l^m(\theta,\phi) \tag{A.14a}$$
$$n = 1, 2, 3, \ldots$$
$$l = n-1, n-2, n-3, \ldots, 0$$
$$|m| = l, l-1, \ldots, 0$$

$$E_n = -\frac{m_0 q^4}{(4\pi\varepsilon_0)^2 2\hbar^2 n^2} \tag{A.14b}$$

where $R_{nl}(r)$ are known as the radial wavefunctions and $Y_l^m(\theta,\phi)$ are the spherical harmonics. Figure A.4b shows sketches of the energy levels versus quantum number n and position r. Although the wavefunction is now defined by three[12] indices, the energy levels only depend on n. The $1/r$ dependence of the potential well for the hydrogen atom leads to the energy levels becoming more closely spaced as n increases (c.f. infinite square well example). Another occurrence that is very typical of systems in 2 and 3 dimensions is that several eigenfunctions have the same value of energy, known as energy *degeneracy*. The value of the ground state energy level E_1 (−13.6 eV) is known as the *ionization energy* of hydrogen and represents the energy required to completely remove the electron from the hydrogen atom.

The explicit solution for the wavefunction of the ground state of the hydrogen atom is

$$\psi_{100} = \frac{1}{\sqrt{\pi}}\left(\frac{1}{a_0}\right)^{3/2} e^{-r/a_0} \tag{A.15a}$$

where a_0 is known as the *Bohr radius*

$$a_0 = \frac{4\pi\varepsilon_0\hbar^2}{m_0 q^2} = 0.529\,\text{Å} \tag{A.15b}$$

which corresponds to the most probable radius for the electron.[13] The Bohr radius is an important physical parameter that sets the typical length scale of atomic systems. Two types of hydrogen eigenfunctions or orbitals that are relevant to electronic devices are illustrated in Fig. A.4c: *s*-orbitals such as ψ_{100} are spherically symmetric while *p*-orbitals (e.g., ψ_{210}) possess a directionality or polarization with a dumbbell-like shape.

[12] This corresponds to the three spatial dimensions of the hydrogen atomic system.

[13] Found from the maximum value of the ground state probability distribution integrated over a spherical shell.

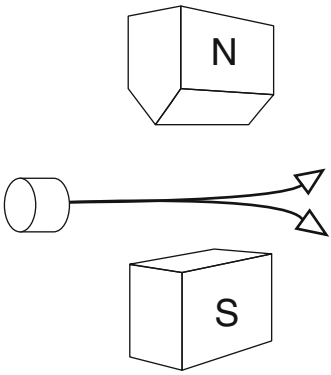


Fig. A.5 Stern–Gerlach experiment and spin. If a beam of electrons is directed towards the nonuniform magnetic field it splits into two parts; half are deflected upwards and half downwards due to the two possible values of spin (Note: the Lorentz force due to the charge on the electrons is ignored in this diagram)

Periodic Table

Two additional concepts are required before we can explain the main features of atomic elements beyond simple hydrogen:

1. Spin

Electrons possess an intrinsic angular momentum known as *spin*, which leads to a magnetic moment.[14] Electron spin can take one of two values (usually referred to as spin up and spin down):

$$\mathbf{s} = \pm\frac{\hbar}{2} \tag{A.16}$$

Thus for every energy eigenfunction which represents an electron there also exist two possible spin states. Therefore in addition to the wavefunction we must also specify the spin value in order to uniquely define the electron state.

Spins were experimentally identified in 1922 via the Stern–Gerlach experiment, a simplified representation of which is shown schematically in Fig. A.5.

2. Pauli exclusion principle

An important physical property of particles with half-integer spin (like electrons) is encapsulated by the following principle:

No more than one electron can be in any given state (wavefunction and spin state) at the same time.

[14] Qualitatively, one can think of the electron wave possessing a circular polarization and thus a localized "current loop" leading to an intrinsic magnetic field. Clockwise and counterclockwise rotation can be related to the two possible values of spin.

A proof of this statement is beyond the scope of this text.[15] This principle plays a crucial role in determining the arrangement and allowed states of electrons in atoms, molecules, and solids. For example, it tells us that each energy level in the infinite square well problem discussed earlier can have at most two electrons (spin up and down). It also explains the strong repulsive interaction force when two pieces of matter are pushed against each other.

Using the results for the hydrogen atom and these two concepts we can now approximate quite well the basic electronic structure for most of the elements in the periodic table of interest for electronic devices.[16]

For example, Si (element 14) has the configuration $1s^2 2s^2 2p^6 3s^2 3p^2$. Here the first digit corresponds to the energy level index, n, while the superscripts denote the number of electrons in each state, where

$$\begin{aligned} l &= 0, 1, 2, 3, \ldots \\ &s, p, d, f, \ldots \end{aligned}$$

Note that the above configuration is for *isolated* Si atoms. When atoms are brought together to form solids and molecules the outer or *valence* electrons almost always reconfigure themselves to achieve the most energetically favorable or stable configuration as will be discussed in the following section. A periodic table is given in Appendix B as reference.

A.2 Semiconductor Physics

A.2.1 Crystal Structure

When atoms are brought together they can form crystalline solids or crystals. The atoms in a crystal are arranged in a *periodic* array which can be described in terms of a *lattice*. Most materials used for electronic devices are usually based on the *cubic* lattices shown in Fig. A.6a. Some standard definitions are used to define planes and directions in a crystal. Most of these are straightforward extensions of ideas from Cartesian geometry. Figure A.6b illustrates the use of Miller indices to define the planes in crystal. In brief, the plane denoted by the triplet (h,k,l) is found by taking the inverse of the point that it cuts on each axis and reducing (if necessary) to the smallest set of three integers. If a plane cuts on the negative side of the origin,

[15] Note that the Pauli exclusion principle only applies to Fermions (e.g., electrons, and most particles with mass). Bosons (e.g., light or photons, which have integer values of spin), on the other hand, can have more than one particle in a given state simultaneously.

[16] For heavier elements (~60 and above) the simple hydrogen atom-based shell-filling model begins to break down as electron–electron interactions and other phenomena such as relativistic effects become more important.

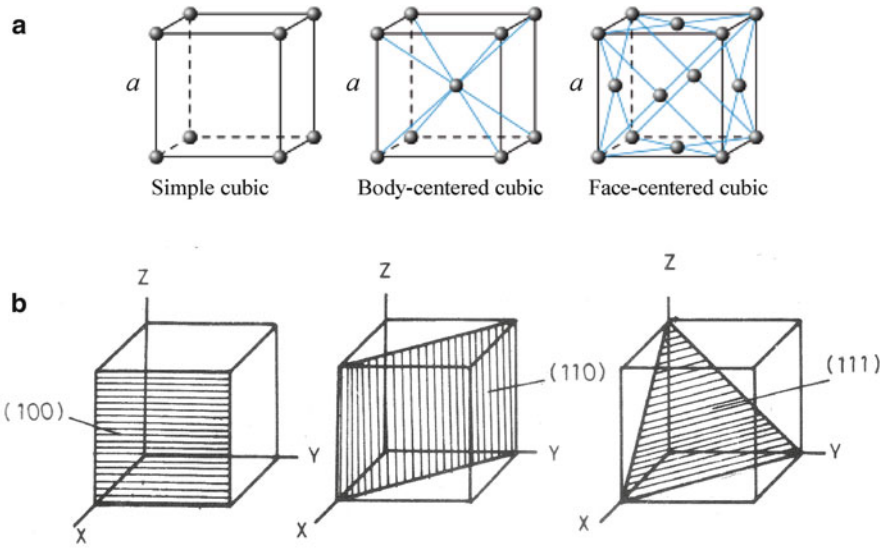


Fig. A.6 (**a**) Cubic lattice crystal structures (*a* denotes the length of the sides in the cubic unit cell). (**b**) Miller indices for some important crystal planes (after W. H. Miller, 1839)

the corresponding index is negative, indicated by placing a minus sign or bar above the index. Directions within a crystal are found in the usual way and denoted by the lattice vector [*h*,*k*,*l*], where the coordinates are multiples of the unit axis vectors.

Two very important crystal structures for electronics are *diamond* and *zincblende*. The diamond structure (Fig. A.7a) consists of two face-centered cubic (fcc) lattices offset from each other by [¼, ¼, ¼]. Silicon and germanium (in addition to carbon, or diamond itself) crystallize in the diamond structure. The zincblende structure is identical to diamond except that the two fcc lattices now contain different atoms. Compound materials such as gallium arsenide and indium phosphide in addition to ternary and quaternary materials (AlGaAs, InGaAs, etc.) and alloys such as SiGe correspond to zincblende crystals.

Bonding Orbitals

In both the diamond and zincblende structures each atom has 4 nearest neighbors, which lead to *tetrahedral* bonding (Fig. A.7b). The atomic bonding exhibited by crystals of Si and Ge is referred to as *covalent bonding*, due to the overlap of valence electron wavefunctions from neighboring atoms. This sharing of electrons results in bonds that are very strong and highly directional.

In order to describe how the covalent bond forms one usually employs the concept of *hybridization* as discussed in the following example.

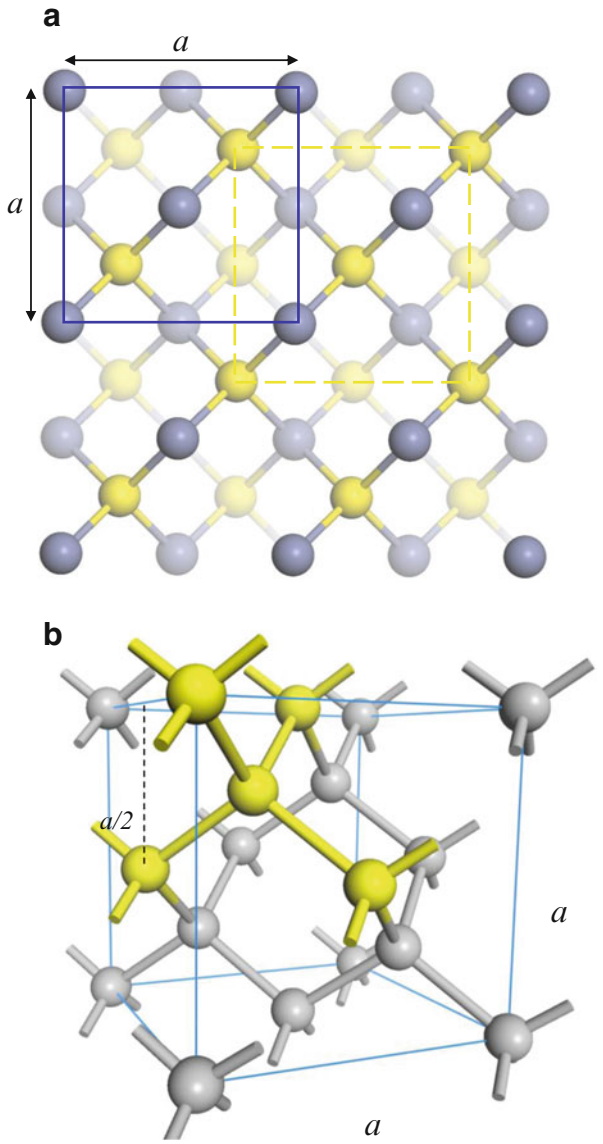
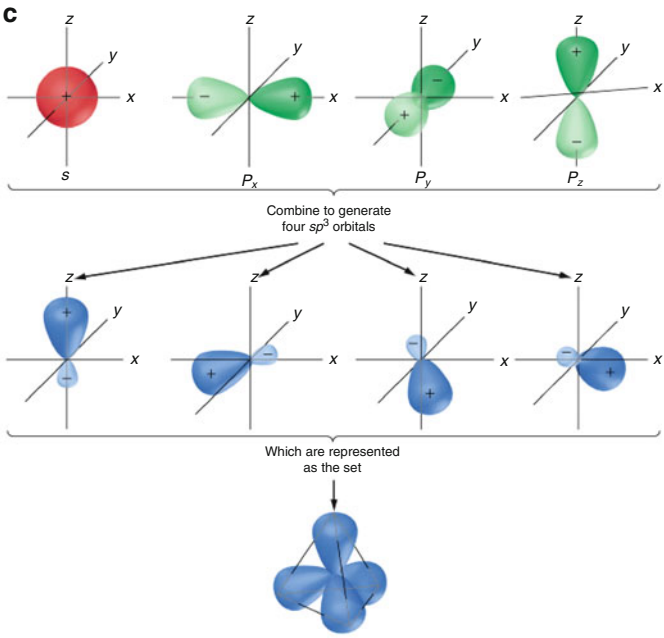


Fig. A.7 (**a**) Diamond and zincblende crystal structure composed of two interpenetrating fcc lattices. Two planes or faces of the fcc cubic unit cells are indicated in the top view shown (blue face in foreground). (**b**) Four nearest neighbors highlighted in the diamond cubic unit cell. (**c**) Illustration of sp^3 hybridization concept leading to tetrahedral bonding. (After R. H. Petrucci, F. G. Herring, J. D. Madura, C. Bissonnette, *General Chemistry: Principles and Modern Applications*, 10th Edition, Prentice-Hall, 2010.) (**d**) Electron microscopy cross-section image of a MOSFET device showing different levels of crystalline order (After S. Monfray et al., Solid-State Electron. **48**, 887 (2004))



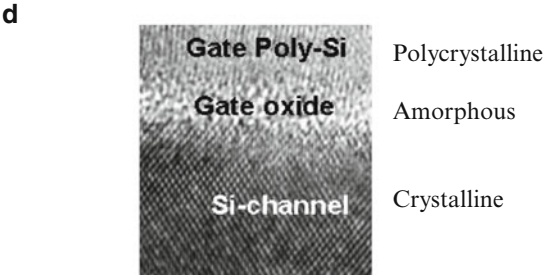


Fig. A.7 (continued)

Example A.5: sp^3 Hybridization in Silicon We saw earlier that the electronic configuration of a silicon atom in its ground state is given by

$$1s^2 2s^2 2p^6 3s^2 3p^2$$

However, when Si atoms combine to form tetrahedral bonds in the diamond configuration the atoms are first promoted to the electronic configuration:

$$1s^2 2s^2 2p^6 3s 3p^3$$

We now have one valence s-electron and three valence p-electrons that can be combined to form four tetrahedral bonds. Hybridizing these $3s$ and $3p$ valence

orbitals is energetically favorable because it increases the wavefunction overlap between atoms, thus creating a more stable bonding structure as illustrated in Fig. A.7c. Note that because of the Pauli exclusion principle the two electrons participating in each bond have opposite or *antiparallel* spins.

The tetrahedral bonding in zincblende structures such as GaAs possesses both a covalent and ionic character (due to charge transfer between the two different atoms).

Although the discussion on crystal structure has assumed a perfect periodic collection of atoms, real systems invariably contain defects and impurities that cause deviations from the ideal crystal structure. Figure A.7d illustrates the varying amounts of disorder in a solid that can occur in practice by examining the cross section of a silicon MOSFET structure from gate to channel (see Chap. 4). An amorphous material has the least amount of long-range order and the atoms can be considered randomized except over very small distances. A polycrystalline material generally contains domains or regions with crystalline order that can vary in size from submicron to many microns. As the domains in a poly-crystal become larger we eventually reach the perfect crystal. Note that even the most perfect crystals will in practice have finite extent and thus the surface of a crystal will always be a source of defects.

A.2.2 Electrons in Crystals; Metals, Insulators, and Semiconductors

There is a very large variation in the electrical conductivity of materials. For example, the resistance of a good metal at room temperature can be up to ~10^{24} times less than that of an insulator. If one considers that most solids contain approximately 10^{23} atoms/cm^3, getting charged particles such as electrons to efficiently move through a material would seem nearly impossible. This difficulty can be reconciled using what is known as the *energy band theory* of solids.

Two methods are commonly used to describe how electrons behave in crystals:

One approach considers how the energy levels of isolated atoms change when they are brought together to form a crystal (Fig. A.8a). The discrete atomic energy levels will begin to split as their wavefunctions begin to overlap (due to the Pauli exclusion principle). At some point the atoms will reach their stable configuration with the original discrete levels now spread out to form an essentially continuous distribution of energies over a certain range.

The other approach examines the quantum mechanical properties of electrons in a crystal by attempting to solve the Schrödinger equation directly for particles traveling in the periodic potential of a crystal lattice. This can be thought of as an extension of the single potential barrier problem examined earlier to an infinite number of regularly space barriers. Figure A.8b shows the resultant energy vs. wave vector dispersion relation for such a system in the case of a model one-dimensional lattice. The result is very similar to the parabolic dispersion relation for a free particle given in Sect. A.1 except that at particular values of wave vector the parabola is no longer continuous but forms gaps in energy.

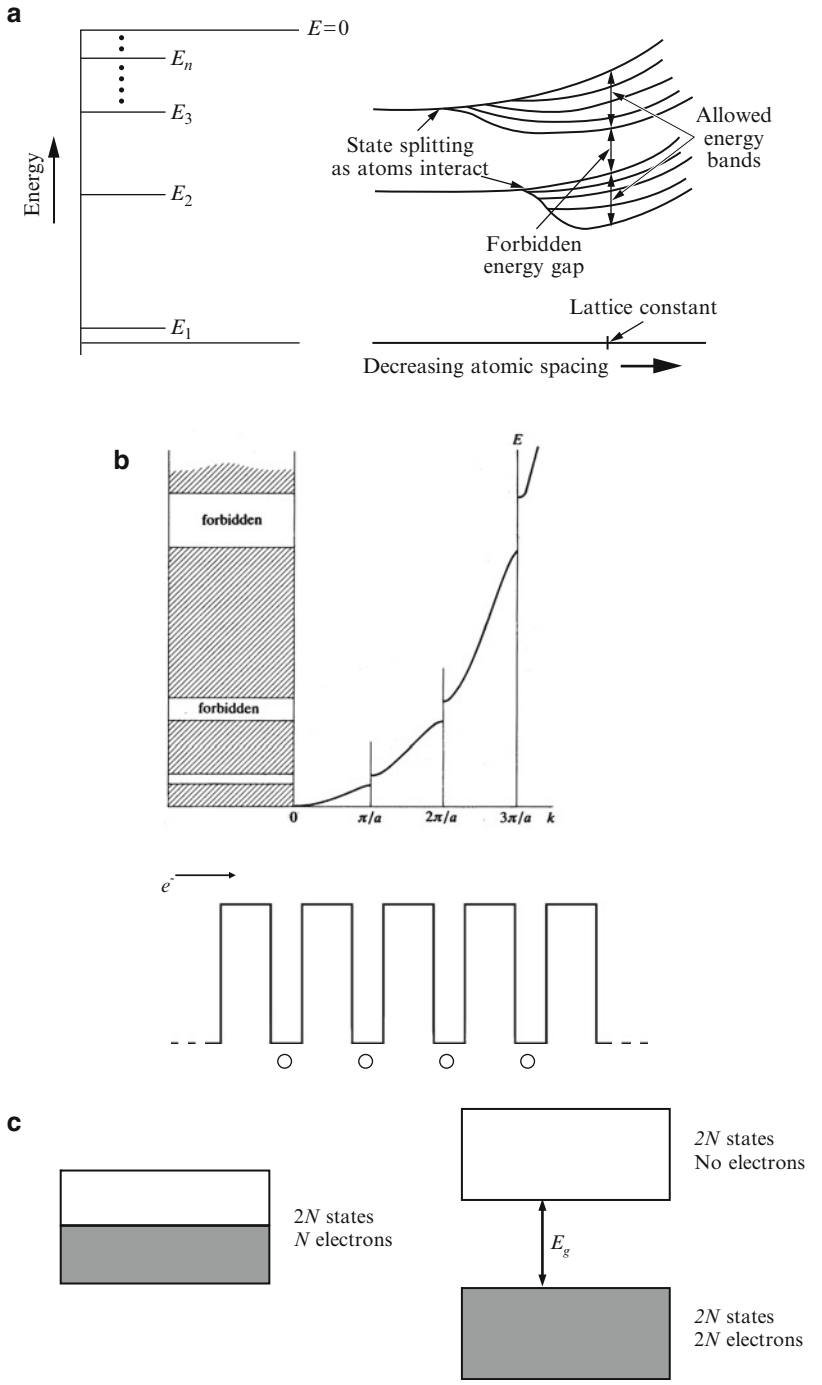


Fig. A.8 Energy band theory. (**a**) Splitting of energy levels as atoms are brought closer together. (Adapted from [6].) (**b**) Energy vs. wave vector dispersion relation for a particle traveling in a one-dimensional periodic potential (a simple model of a crystal with the atoms represented by square potential wells is shown (Kronig–Penney model)). (**c**) Metal vs. insulator according to band theory. In the metal (*left*) the uppermost filled band is only partially filled, which allows charge

A similar phenomenon arises whenever a wavelike disturbance is incident on a periodic potential, i.e., *Bragg reflection*. At certain wavelengths there will be strong constructive interference of the multiple wavefronts being reflected by a periodic system. The Bragg condition states that strong reflection will occur when

$$n\lambda = 2a\sin\theta \tag{A.17}$$

where n is a positive integer and a corresponds to the lattice spacing for a wave incident at an angle θ. For a one-dimensional lattice the incident angle is 90° and the Bragg condition corresponds to $n\lambda = 2a$ or, in terms of wave vector, $k = n\pi/a$. At these special values of wave number electrons are strongly reflected and cannot propagate through the crystal. Away from these values, however, electrons in a crystal can behave very much like free particles.

In practice, a combination of approaches and significant computational effort are usually needed to explain the electronic properties of real materials. However, regardless of the particular details, the formation of allowed and forbidden electron *energy bands* in crystalline solids is predicted, which provide an avenue for electrons to travel through the crystal.[17]

Band theory allows a straightforward explanation of the difference between a metal and an insulator[18]: Each band has $2N$ available electron states where N is the number of unit cells making up the crystal (the factor of 2 accounts for spin). How the bands are filled with electrons determines whether a material is a metal or an insulator. Referring to Fig. A.8c, if the uppermost populated energy band is only partially filled (say, half-filled) then there are plenty of higher energy states available for electrons to gain kinetic energy so that they can contribute to current flow and thus the crystal is able to conduct electricity. If, on the other hand, the uppermost band is completely filled then there is no easy and continuous way for charge carriers to gain energy because there is a forbidden *energy band gap* E_g before the next band of states becomes accessible. The crystal is therefore insulating and cannot conduct electrical current.[19]

Fig. A.8 (continued) carriers to gain energy and contribute to current flow. In the insulator (*right*) the band is completely filled and separated from the next available states by the band gap energy, which means there is no continuous way to increase the energy of charge carriers and thus no current can flow

[17] There are some important exceptions where the band theory of solids is not valid, such as materials that have very strong electron–electron interactions. However, the vast majority of crystalline solids are very accurately described using band theory.

[18] These arguments were first put forth by Alan Wilson in 1930.

[19] Note that at some point an insulator subjected to very large external forces will conduct electricity via various breakdown mechanisms.

For example, diamond, silicon, and germanium crystals each contribute an even number of valence electrons per unit cell so their bands are completely filled up to the so-called *valence band* and they are insulating at 0 K. *Semiconductors* are usually defined as insulators with a band gap that is typically in the range of a few eV or less. The band gap energy is an extremely important parameter when dealing with semiconductors and semiconductor devices.

The band structures of actual three-dimensional crystals are quite complicated as illustrated for the two semiconductors shown in Fig. A.9a. Two important concepts can be deduced from these examples:

The first is direct vs. indirect band gaps—for a *direct band gap* semiconductor (e.g., GaAs) the energy band gap occurs at the same value of wave vector (or crystal momentum); otherwise we have an *indirect band gap* material (e.g., Si). The type of band gap is particularly important for optoelectronic devices because it affects the probability of light absorption and emission in a semiconductor, with direct band gap materials being much more efficient for both processes, especially light emission.

The other notion is that of *effective mass*: near the top and bottom of bands they can be approximated by parabolas, similar to a free particle. Thus one can define the effective mass of electrons in a material as

$$m^* = \frac{\hbar^2}{d^2E/dk^2} \tag{A.18}$$

The effective mass is inversely proportional to the curvature of the band—if the curvature is large the effective mass is small and vice versa. The effective mass approximation allows many of the complicated band structure details to be described by a single parameter.

Experimental parameters for some of the standard semiconductors are listed in Appendix B.

A.2.3 Electrons and Holes; Doping

When dealing with electronic devices it is convenient and useful to simplify the band structure of a material by only considering the band energies vs. position instead of the full energy vs. momentum band diagram. Figure A.9b illustrates a standard *band edge diagram* for a semiconductor, where the valence band edge is denoted by E_v and the next higher-lying band edge is labeled by E_c for what is termed the *conduction band*. As discussed above, the two band edges are separated by the band gap E_g. Once again referring to the figure, consider what happens when electrons are excited from the valence band to the conduction band of a semiconductor: The electrons promoted into the conduction band can now participate in electronic transport. In addition, the electrons in the valence band now have some empty states available for them to also participate in current flow. However, instead of keeping track of all these electrons, the vacant states that are left in the otherwise full valence band can be treated as if

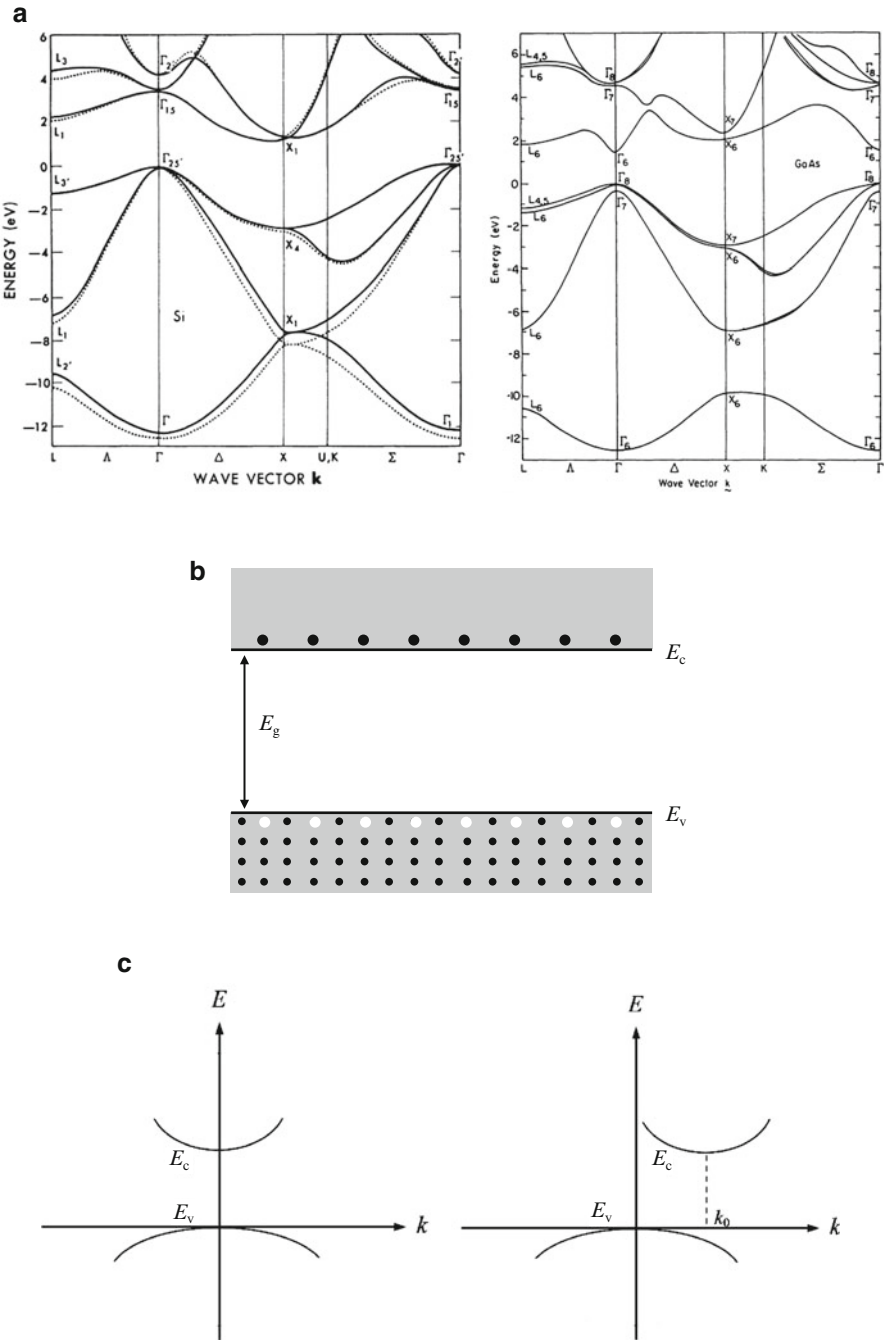


Fig. A.9 (**a**) Band structures for Si and GaAs along high-symmetry directions. All the bands are usually plotted about the origin as shown without loss of generality (cf. Fig. A.8b). (After M. L. Cohen, J. R. Chelikowsky, *Electronic Structure and Optical Properties of Semiconductors*, 2nd Edition, Spring-Verlag, 1988.) (**b**) Energy band edge diagram for a semiconductor. Electron–hole pairs are created when carriers are excited from the valence band to the conduction band. (**c**) Illustration of direct (*left*) vs. indirect (*right*) band gap semiconductor

they were particles called *holes*.[20] A hole acts like a particle with positive charge q. Both electrons and holes contribute to electric current.

Notice that the electron effective mass is negative at the top of the valence band. This is equivalent to a hole with positive effective mass and positive charge. Thus we use the notation

$$m_p = -m_n \tag{A.19a}$$

for the hole effective mass at the top of the valence band and

$$m_n \tag{A.19b}$$

for the electron effective mass at the bottom of the conduction band. As a result we can write

$$E = E_c + \frac{\hbar^2(k-k_0)^2}{2m_n}; E = E_v - \frac{\hbar^2\left(k-k'_0\right)^2}{2m_p} \tag{A.20}$$

for the electron energies in the conduction band and valence band, respectively. Note that if $k_0 = k'_0$ the band gap is direct; otherwise it is indirect (see Fig. A.9c).

Donors and Acceptors

In an *intrinsic* material the number of electrons in the conduction band is equal to the number of holes in the valence band. At a given temperature, such *electron–hole pairs* are excited across the band gap in a semiconductor by thermal energy. In terms of carrier concentration per unit volume this is expressed as

$$n = p = n_i \tag{A.21}$$

where n and p are the electron and hole concentrations, respectively, and n_i is called the *intrinsic carrier concentration*.

On the other hand, in an *extrinsic* material one type of carrier has a greater concentration (*majority* carriers) than the other (*minority* carriers). This is achieved by substitutional *doping* of the crystal with impurity atoms that either *donate* electrons to the conduction band or *accept* electrons from the valence band. When electrons are the dominant carrier type in a semiconductor it is called *n-type*.

[20] A very often used analogy considers the electrons as a "sea" and the holes as "bubbles" in this sea. Even though the bubbles do not contain any water they can still be treated as particles as long as one remembers that it is actually the water that moves in the opposite direction. The same principle applies to the electrons in the valence band, which move oppositely to the holes. Holes can also be said to "float" to the top of the valence band in order to reach their lowest energy.

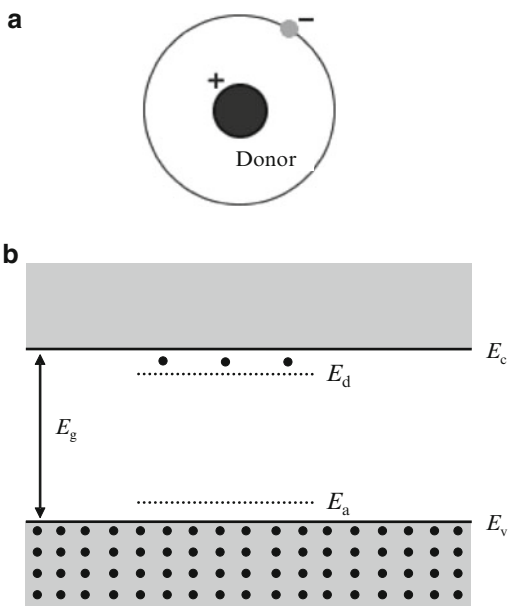


Fig. A.10 Impurity dopant atom binding energies. (**a**) Donor impurity hydrogenic model. (**b**) Band edge diagram showing impurity energy levels and carrier distribution at low temperature

Similarly, when holes are the dominant carriers the material is called *p-type*. Typical donor atoms for semiconductors are from column V of the periodic table, e.g., P, As, Sb. Acceptors commonly used are from column III such as B, Al, Ga, and In.

Hydrogenic Model for Impurity Binding Energy

Recall for the hydrogen atom the ionization energy was given by

$$E = -\frac{m_0 q^4}{(4\pi\varepsilon_0)^2 2\hbar^2}$$

which turned out to be –13.6 eV. In other words the energy required to remove an electron from a hydrogen atom, or the *binding energy*, is 13.6 eV.

A donor impurity atom in a semiconductor can be approximated as a hydrogen atom by considering the outer electron as being loosely bound (Fig. A.10a). We can now use the same results as for the hydrogen atom by substituting the effective mass and the dielectric constant of the semiconductor (or relative permittivity, ε_r) in order to approximate the binding energy of a donor impurity atom:

$$E = \frac{m_n q^4}{(4\pi\varepsilon_0\varepsilon_r)^2 2\hbar^2} = \frac{13.6}{\varepsilon_r^2}\frac{m_n}{m_0}\text{eV} \tag{A.22}$$

The same thing can be done for acceptor impurities by simply replacing the electron effective mass with the hole effective mass.

Measured values of the donor binding energies in most semiconductors are typically between 10 and 50 meV and similarly for acceptors. This means that the donor and acceptor impurity levels are quite close to either the conduction or valence band edges (Fig. A.10b), which is important if doping is to be effective around room temperature.[21]

A.2.4 Carrier Concentration in Thermal Equilibrium[22]

A fundamental result from statistical mechanics is the *Fermi–Dirac distribution*, which states the probability that an electron will occupy a state with energy E at temperature T as

$$f(E) = \frac{1}{1 + e^{(E-E_F)/k_B T}} \tag{A.23}$$

where k_B is Boltzmann's constant and E_F is the *Fermi level*. It is useful to think of the Fermi level, also known as the chemical potential, as representing the surface level of the "sea" of electrons in a material.[23] A plot of the Fermi–Dirac distribution is shown in Fig. A.11a. Note that $f(E)$ is always equal to ½ when the energy E is equal to E_F. In addition, the *hole* probability distribution is given by $1 - f(E)$, since a hole is the absence of an electron.

Combining the Fermi–Dirac distribution with the band edge diagram of a semiconductor results in Fig. A.11b. It can be seen that the Fermi level of a semiconductor moves closer to the conduction band edge for an n-type material (more electrons) whereas it is lowered toward the valence band edge for p-type doping (less electrons). For an intrinsic semiconductor with equal numbers or electrons and holes, the Fermi level lies very close to the middle of the band gap (see discussion below).

Before continuing, we mention some standard definitions involving the Fermi level: The vacuum energy level (E_0) is a convenient reference in band edge diagrams, which corresponds to the energy of an electron that has been just freed

[21] In other words, the dopants will be easily activated by thermal energy.

[22] A system is considered to be in a state of *thermal equilibrium* if its temperature is equal to that of the surrounding environment (or thermal energy "reservoir") and its properties will not change as long as the temperature remains constant. This implies that the system is not subjected to any external forces or sources of energy that can perturb it from being in equilibrium with its environment (also known as thermodynamic equilibrium).

[23] A higher Fermi level means more carriers in the electron sea and a lower Fermi level means less. Most of the important electronic effects in materials occur in and around the Fermi level (somewhat similar to the disturbances that occur on the surface of a body of water).

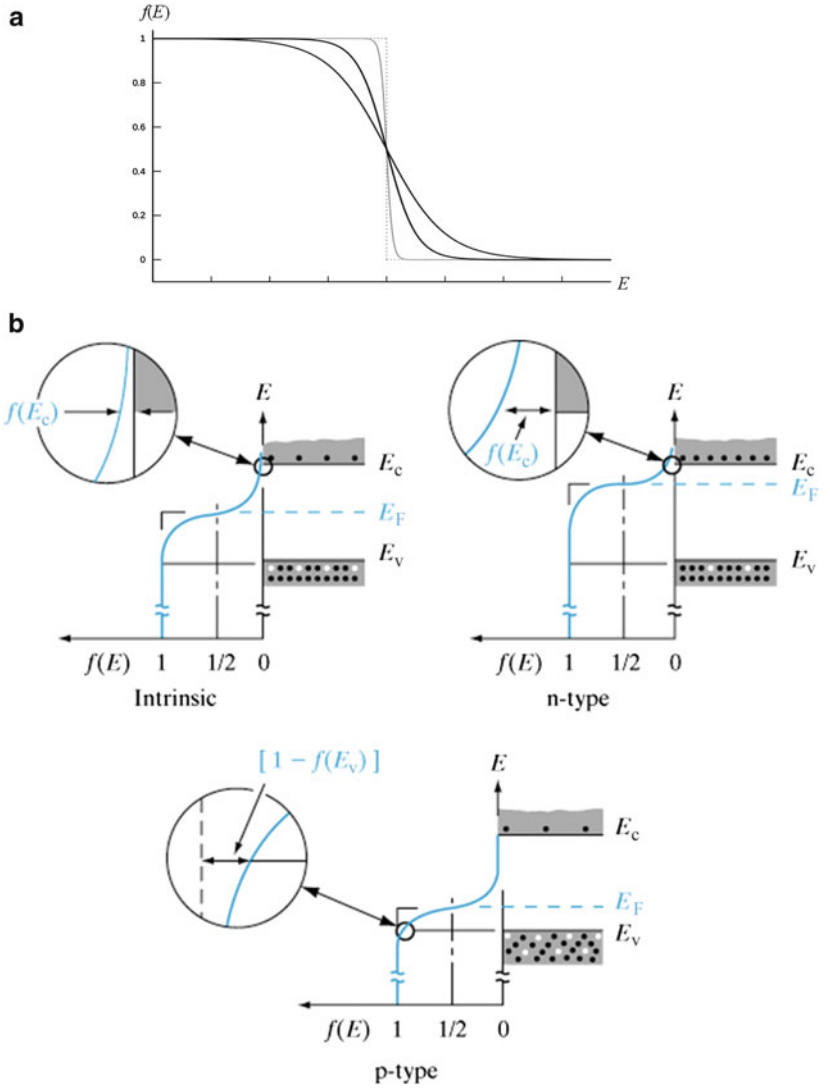


Fig. A.11 (**a**) Fermi–Dirac Distribution at different temperatures. *Dotted curve* is for a temperature of 0 K and the curves gradually spread out as temperature is increased. Note that this distribution is for a system in *thermal equilibrium*. (**b**) Fermi–Dirac distribution superimposed on semiconductor band edge diagram (After B. Streetman, S. Banerjee, *Solid State Electronic Devices*, 6th Edition, Prentice-Hall, 2005)

from a material. The difference between the Fermi level and E_0 is known as the *work function*, $q\Phi$, of the material (similar to the binding energy of an atom). Since the work function will vary with doping level in a semiconductor, we also define the difference between the conduction band edge, E_c, and the vacuum level as the *electron affinity*, $q\chi$, which is constant for a given semiconductor.

Calculation of Carrier Concentrations

The thermal equilibrium concentration of electrons in the conduction band, n, can be found using

$$n = \int_{E_c}^{\infty} f(E)D(E)dE \tag{A.24}$$

where $D(E)$ is the *electronic density of states* per unit energy (per unit volume) and thus $D(E)dE$ gives the number of states between E and $E + dE$.[24] To evaluate the integral we can usually assume that the Fermi–Dirac distribution in the conduction band can be approximated by its high-energy exponential tail,

$$f(E) \sim \mathrm{e}^{-(E-E_F)/k_B T} \tag{A.25}$$

known as the Maxwell–Boltzmann distribution. This limit is valid when $E - E_F \gg k_B T$.[25] With this approximation, the following results for the electron and hole carrier concentrations can be obtained:

$$n = N_c e^{-(E_c - E_F)/k_B T}; \quad p = N_v e^{-(E_F - E_v)/k_B T} \tag{A.26a}$$

where

$$N_c = 2\left(\frac{m_n k_B T}{2\pi\hbar^2}\right)^{3/2}; \quad N_v = 2\left(\frac{m_p k_B T}{2\pi\hbar^2}\right)^{3/2} \tag{A.26b}$$

are known as the *effective density of states* in the conduction and valence band, respectively. Note that the product of electron and hole concentrations in a semiconductor is constant for a given temperature (mass-action law[26]), which can be written

[24] The density of states per unit volume of a 3D (or bulk) solid is given by

$$D^{3D}(E) = \frac{1}{2\pi^2} \cdot \left(\frac{2m^*}{\hbar^2}\right)^{3/2} \cdot E^{1/2}$$

It is important to multiply the probability by this factor when calculating the carrier density via Eq. (A.24) since the probability distribution on its own does not contain full information (for example, the Fermi–Dirac distribution is finite in the band gap, although there are no available states there). The rapidly decaying exponential of the Maxwell–Boltzmann distribution in the conduction band leads to the majority of carriers being clustered near the band edge (similar comments apply to holes in the valence band).

[25] In practice the Maxwell–Boltzmann distribution is a good approximation as long as the Fermi level is 2 or 3 times $k_B T$ away from the band edges.

[26] This important and very useful result is an expression of thermal equilibrium and applies whether or not the semiconductor is doped, with the only assumption being that Maxwell–Boltzmann statistics are valid.

$$np = {n_\mathrm{i}}^2 \tag{A.27}$$

where

$$n_\mathrm{i} = \sqrt{N_\mathrm{c} N_\mathrm{v}} e^{-E_\mathrm{g}/2k_\mathrm{B}T} \tag{A.28}$$

is the intrinsic carrier concentration. This shows mathematically that the intrinsic carrier concentration *increases with increasing temperature* and *decreases with increasing band gap*, as expected for the thermal excitation of electron–hole pairs.

Finally, we can also express the equilibrium electron and hole concentrations in terms of the intrinsic concentration and Fermi level as

$$\begin{aligned} n &= n_\mathrm{i} e^{(E_\mathrm{F}-E_\mathrm{i})/k_\mathrm{B}T} \\ p &= n_\mathrm{i} e^{(E_\mathrm{i}-E_\mathrm{F})/k_\mathrm{B}T} \end{aligned} \tag{A.29}$$

where E_i is the *intrinsic Fermi level*. This form of the equations is useful in practice because n_i is usually well known from experiment for most semiconductors. The position of E_i can be found by equating Eqs. (A.26a) to give

$$E_\mathrm{i} = \frac{E_\mathrm{c} + E_\mathrm{v}}{2} + \frac{3}{4} k_\mathrm{B} T \ \ln\left(\frac{m_p}{m_n}\right) \tag{A.30}$$

This shows that the intrinsic Fermi level lies very close to the center of the band gap, with only a slight offset due to the difference in electron and hole effective mass.

Space-Charge Neutrality

If we start with an electrically neutral semiconductor in thermal equilibrium it must remain neutral regardless of the number of holes, electrons, or ionized impurities present. This condition is known as (global) *space-charge neutrality* and can be expressed as

$$n + N_\mathrm{a}^- = p + N_\mathrm{d}^+ \ \text{ or } \ n = p + \left(N_\mathrm{d}^+ - N_\mathrm{a}^-\right) \tag{A.31}$$

where N_a^- and N_d^+ represent the concentration of ionized dopant atoms. Equation (A.31) and the mass-action law [Eq. (A.27)] are the two requirements that must be maintained at equilibrium. Assuming all the dopant atoms are ionized[27] and substituting

$$p = n_\mathrm{i}^2 / n$$

into Eq. (A.31), we obtain

[27] We will normally assume in this text that all dopants are ionized and therefore not explicitly include the superscripts in the symbols for dopant concentration.

$$n - \frac{n_i^2}{n} = N_d - N_a \Rightarrow n = \frac{N_d - N_a}{2} + \left[\left(\frac{N_d - N_a}{2}\right)^2 + n_i^2\right]^{1/2} \tag{A.32a}$$

for the electron concentration in a semiconductor (keeping only the positive root), and the analogous expression for holes is given by

$$p = \frac{N_a - N_d}{2} + \left[\left(\frac{N_a - N_d}{2}\right)^2 + n_i^2\right]^{1/2} \tag{A.32b}$$

Example A.6: Carrier Concentration Calculation Find the electron and hole concentrations and location of E_F for silicon at 300 K doped with 3×10^{16} cm^{-3} arsenic atoms and 2.9×10^{16} cm^{-3} boron atoms.

Using Eq. (A.32a) we can see that since $N_d - N_a >> n_i$ (see Appendix B) that $\Rightarrow n \approx N_d - N_a = 10^{15}\text{cm}^{-3}$ and therefore $p = \frac{n_i^2}{n} = 2.25 \times 10^5\text{cm}^{-3}$. The position of the Fermi level can now be found using Eq. (A.29) for the electron concentration:

$$n = n_i e^{(E_F - E_i)/k_B T};$$
$$\Rightarrow E_F - E_i = k_B T \ \ln\left(\frac{n}{n_i}\right) = 0.289 \ \text{eV}$$

A.2.5 Equilibrium Fermi Level and Band Edge Diagrams

Recall that in thermal equilibrium there can be no external forces or excitations acting on the material or device under consideration. In other words, there is a detailed balance between the various physical processes occurring in the system. From the perspective of electronic devices this implies the following statement holds

*In **thermal equilibrium** there is no net current flow and thus the **Fermi level must be constant** throughout the material or device.*

As a function of distance x, this condition requires

$$\frac{dE_F}{dx} = 0 \tag{A.33}$$

in thermal equilibrium. As illustrated in Fig. A.12, regardless of the particular material or device details the Fermi level must be constant in order to satisfy the condition of zero net current flow and this is represented by a horizontal straight line in the thermal equilibrium band edge diagram.

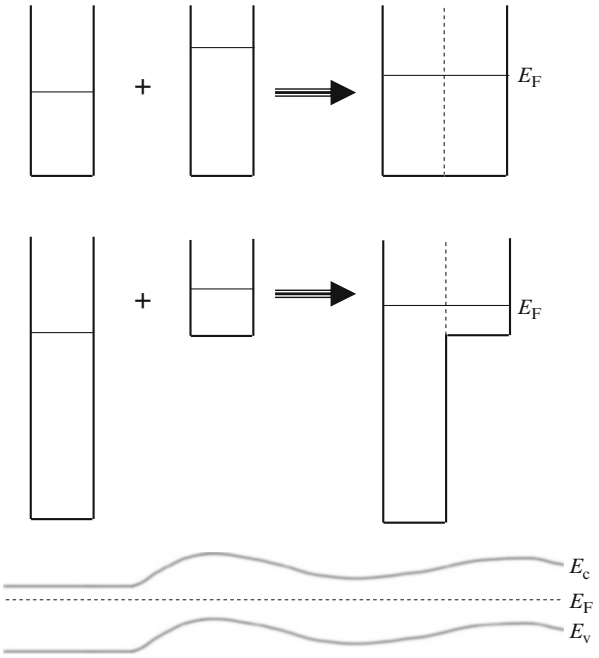


Fig. A.12 Thermal equilibrium Fermi level. The water analogy shows that the level of liquid will always be constant in equilibrium regardless of the amount of water in the vessels before they are combined. Water will flow from high level to low level in order to establish thermal equilibrium as shown in the two examples. The same ideas apply to the electrons in a thermal equilibrium band edge diagram of a material or device regardless of the doping levels or type of material, etc.

A.2.6 Electronic Transport

At a given temperature, electrons and holes in a semiconductor possess average thermal energy given approximately by[28]

$$\frac{1}{2}m^* v_{th}^2 = \frac{3}{2}k_B T \tag{A.34}$$

where v_{th} is the thermal velocity. This causes random motion of the charge carriers resulting in collisions or scattering with the lattice impurities and between carriers etc. In thermal equilibrium we have seen that this type of random motion does not produce any net current flow.

Drift Current

On the other hand, when an external electric field is applied to the semiconductor the carriers acquire a *drift velocity*: Figure A.13a shows the combined effects of the

[28] This result comes from the Equipartition theorem of classical statistical mechanics, which states that for a system in thermal equilibrium particles have an average kinetic energy of 1/2 $k_B T$ per spatial degree of freedom.

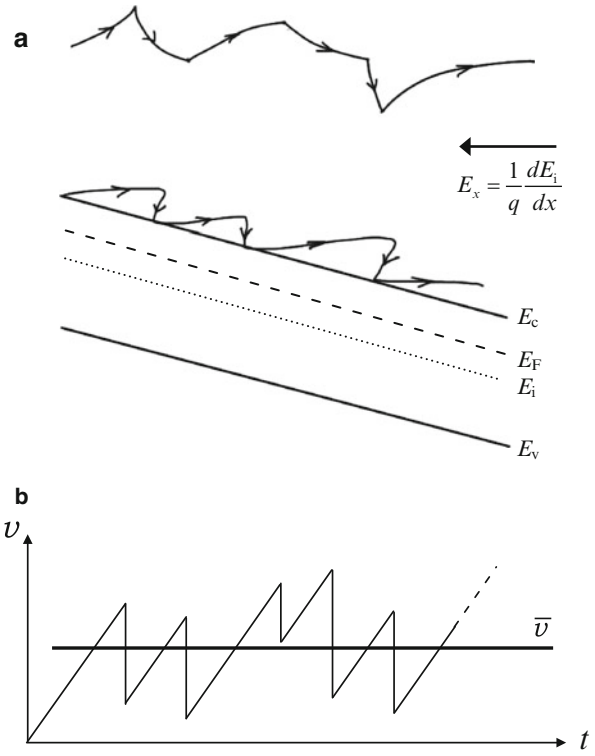


Fig. A.13 (**a**) Illustration of carrier drift in an applied electric field (top—real space, bottom—band edge diagram). The electric field is found from the gradient of the potential, which can be defined in terms of either of the parallel band edges or the Fermi level. A convenient definition is in terms of the intrinsic Fermi level as indicated on the diagram (the electrons drift towards the field). (**b**) Velocity versus time for a carrier undergoing drift transport leading to an average drift velocity

applied electric field and random scattering processes on carrier motion that leads to drift transport in a semiconductor. Figure A.13b is a sketch of the corresponding carrier velocity vs. time in a constant applied field showing the initial increase in velocity as carriers are accelerated by the field and subsequent loss of energy due to the various scattering processes. The time interval between scattering events is given by the *mean free time*, τ. The mean free time gives the characteristic timescale over which the carriers are accelerated.

If a constant electric field, E_x, is applied in the x-direction the drift velocity acquired by electrons is given by

$$\overline{v} = -\frac{qE_x\tau}{m_n} \tag{A.35}$$[29]

The electron *mobility* is defined as

[29] This result can be obtained by integrating Newton's second law to find the velocity for a force of qE_x and using the mean free time in the resulting expression. Such approaches, which combine aspects of classical and quantum mechanics, are usually referred to as "semi-classical" treatments.

$$\mu_n = \frac{q\tau}{m_n} \tag{A.36}$$

and describes how easily an electron can move in response to an applied electric field. Analogous arguments apply for holes (in general, the drift velocity can be found using $|\overline{v}|=\mu E_x$). Remembering that holes and electrons move in opposite directions in an electric field and thus both contribute positively to the drift current, the general expression for electrical current density, $J = qnv$,[30] can be now be used to write

$$J_x = J_n + J_p = \left(nq\mu_n + pq\mu_p\right)E_x \tag{A.37}$$

as the *total drift current density* due to both electrons and holes in a semiconductor.

Equation (A.37) has the form of *Ohm's law*

$$J_x = \sigma E_x \tag{A.38}$$

with the conductivity, σ, given by the term in brackets,[31,32] We can relate this expression to the resistance, R, by considering a bar of material with cross-sectional area A and length L. In this case the current density and electric field are given by

$$J = \frac{I}{A} \quad \text{and} \quad E = \frac{V}{L} \tag{A.39}$$

respectively, where I is the current and V is the voltage. Substituting into Eq. (A.38) and rearranging terms gives

$$V = \left(\frac{L}{\sigma A}\right)I = \left(\frac{\rho L}{A}\right)I = IR \tag{A.40}$$

which gives Ohm's law in the familiar form relating voltage and current and shows the dependence of resistance on material geometry (the resistivity $\rho = 1/\sigma$ has also been introduced).

[30] Here n simply denotes a generic carrier concentration and v the carrier velocity that lead to a certain flux density of particles per unit area.

[31] The analysis leading to Eq. (A.37) is generally referred to as the Drude model. This model works well for semiconductors but must be modified for metals to take into account that carriers near the middle of a band have energies on the order of E_F and thus the full quantum Fermi–Dirac statistics must be used. The modified approach is known as Drude–Sommerfeld theory.

[32] In general the conductivity will be given by a tensor; however, a one-dimensional treatment usually suffices for first-order device analysis.

Scattering

The two main types of scattering that influence the mobility of charge carriers in a semiconductor crystal are *lattice* (phonon) scattering and *impurity* (defect) scattering.

Lattice scattering *increases* with temperature as the vibrations of the lattice become greater. This type of scattering causes the mobility to scale as a power law with temperature, T^{-n}, with n typically ~1.5–3.

Impurity or defect scattering, on the other hand, *decreases* as temperature increases. This is due to the increased thermal velocity of carriers, which makes them less susceptible to interaction with impurities/defects. The opposing tendencies for the two types of scattering lead to semiconductor carrier mobility that often resembles an inverted arch shape versus temperature.

Scattering data for silicon is given in Appendix B.

Drift Velocity Saturation

The expressions for drift velocity given above predict that the carrier velocity will continue to increase indefinitely as the electric field is increased. In reality, these expressions are only valid for relatively low fields. At very high electric fields the drift velocity becomes comparable to the thermal velocity. Charge carriers are referred to as *hot carriers* when their kinetic energy is greater than the thermal energy of the rest of the crystal. The drift velocity of such carriers approaches a saturation value at high fields and thus the mobility decreases from its low-field value. This is caused by increased scattering at high fields which limits further increase in the drift velocity. In silicon the saturation velocity for both electrons and holes is approximately 10^7 cm/s or about 0.1 % of the speed of light. Carrier velocity vs. electric field data for silicon are given in Appendix B.

Diffusion Current

In addition to the drift of carriers in an electric field another basic mechanism which can contribute to the electrical current in a material is *diffusion*. Diffusion occurs when there is a carrier *concentration gradient*. This causes charge carriers to flow from a region of high concentration to low concentration resulting in a net current. This is similar to diffusive phenomena that occur in chemical systems such as gases and liquids, in addition to the thermal diffusion of heat.

At a finite temperature, the thermal motion of carriers leads to a net flux of current that will depend on the difference in concentration between adjacent regions (i.e., the concentration gradient) as the schematic of Fig. A.14a illustrates.

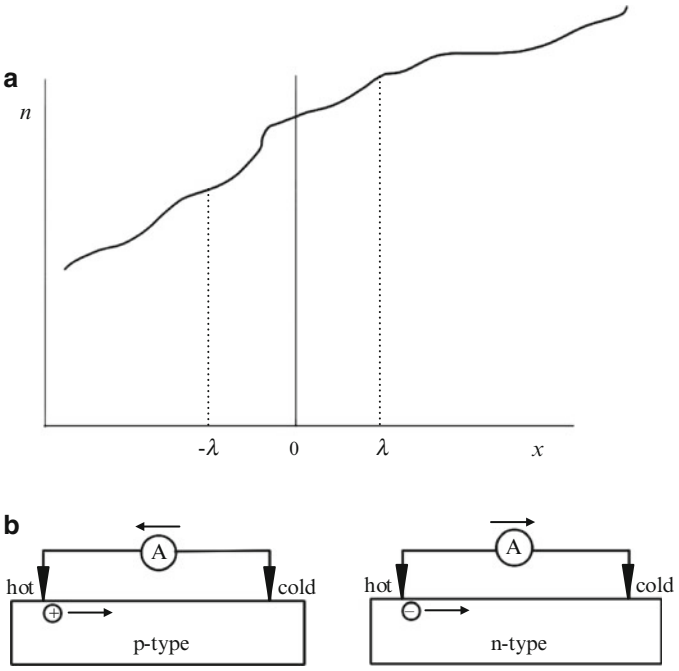


Fig. A.14 Diffusion transport of carriers. (**a**) Schematic displaying a varying electron concentration as a function of position. Carriers within a mean free path λ on either side of the plane at the origin will cross the plane due to random thermal motion at a finite temperature. If the concentration of carriers is not constant this thermal motion will lead to net carrier flow from an area of high concentration to low concentration (towards the left in the diagram). (**b**) Hot probe measurement. In this case, a difference in the thermal velocities of the two regions near the hot and cold probe pins has the same effect as a concentration gradient (in other words there is a velocity gradient instead) and leads to diffusion of carriers from the hot probe to cold

Fick's law, which governs the diffusive flux of carriers, allows the electron diffusion current density to be expressed as

$$J_n = qD_n \frac{dn}{dx} \tag{A.41}$$

where the *diffusion coefficient* (or constant) D_n can be found using

$$D_n = \frac{k_B T}{q} \mu_n \tag{A.42}$$

which is known as the *Einstein relation*.

Analogous results apply for the diffusion of holes due to a concentration gradient.

A *hot-probe* measurement is a convenient technique that relies on diffusion currents to determine the conductivity type of a semiconductor sample. The hot-probe setup consists of two probes, one of which is heated, and an ammeter (Fig. A.14b). No voltage is applied but a current flows when the probes touch the semiconductor whose direction depends on whether the material is n-type or p-type.

We can now write the *total* electron and hole current densities including both drift and diffusion:

$$\begin{aligned} J_n &= q\mu_n n E_x + qD_n \frac{dn}{dx} \\ J_p &= q\mu_p p E_x - qD_p \frac{dp}{dx} \end{aligned} \tag{A.43}$$[33]

Note that because diffusion depends on the gradient of carrier concentration minority carriers can still make a significant contribution to the overall current density even if their total concentration is low.

The drift–diffusion transport equations (A.43) are the typical starting point for standard semiconductor electronic device analysis. (Along with the basic equations of electrostatics and the continuity equations).

The Hall Effect

The total force on a charged particle moving in electric and magnetic fields is given by the *Lorentz force*:

$$\vec{F} = q\left[\vec{E} + \vec{v} \times \vec{B}\right] \tag{A.44}$$

The Hall effect[34] is a direct consequence of this equation and has become a very useful and important experimental technique for the study of semiconductors.

Consider a sample placed in a magnetic field while current is passed through it as shown in Fig. A.15a: Due to the Lorentz force an induced transverse electric field E_H, known as the Hall field, is set up, which produces a Hall voltage V_H that can be measured with external contacts as shown. The force created by the Hall field opposes the Lorentz force, which in steady state leads to a balance that can be expressed mathematically as

$$qE_H = q\overline{v}B \tag{A.45a}$$

[33] Note that the negative sign for the hole current density expression arises from the fact that both electrons and holes diffuse in the same direction (from high to low concentration) but have opposite charges.

[34] First observed by Hall for metals in 1879.

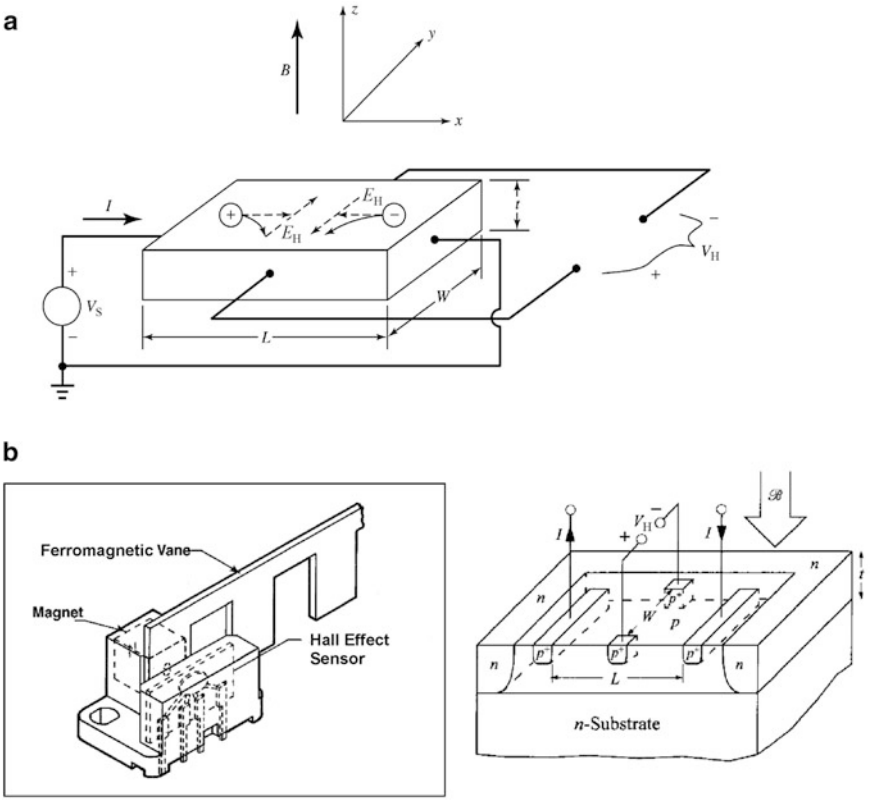


Fig. A.15 Hall effect. (**a**) Experimental setup showing movement of electrons/holes under the influence of an applied magnetic field in the z-direction due to the Lorentz force interaction. (After [6]) (**b**) Hall effect linear motion sensor example and integrated circuit device structure schematic (After S. M. Sze, K. K. Ng, *Physics of Semiconductor Devices*, 3rd Edition, Wiley Interscience, 2007)

$$\Rightarrow E_H = \frac{J_x B}{qp} \text{ for holes;} \quad E_H = -\frac{J_x B}{qn} \text{ for electrons} \tag{A.45b}$$

These equations can be rewritten as

$$E_H = R_H J_x B \tag{A.45c}$$

where R_H is the Hall coefficient, which can be used to determine the majority carrier type (based on its sign), concentration, and mobility of a semiconductor.[35]

[35] Our derivation has implicitly assumed that only one type of carrier is important in the semiconductor, in other words the doping level is such that the minority carriers do not need to be considered. (If this is not the case and both electrons and holes are present in comparable quantities the Hall coefficient expressions become fairly complex.) In addition, note that a numerical factor of order unity usually multiplies R_H to obtain quantitative agreement with experiment.

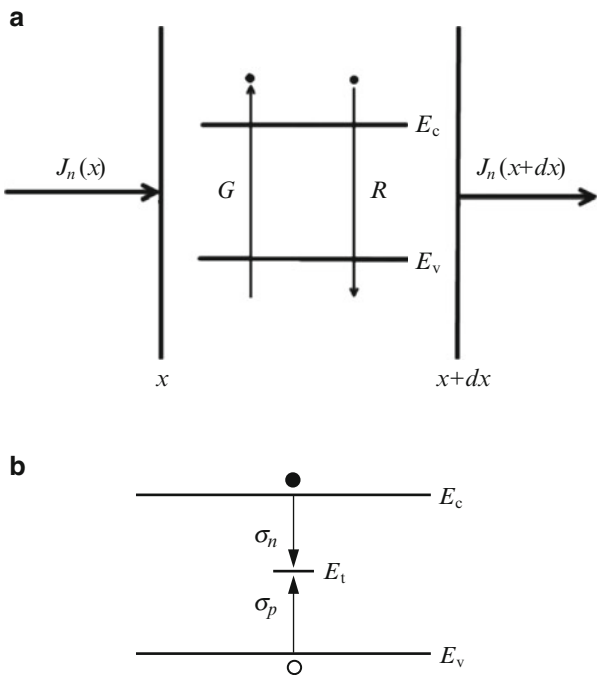


Fig. A.16 (**a**) Conservation of charge in a thin slice of semiconductor used to derive continuity equations. The diagram shows the rate of change of electrons can only be due to current flow into a region or generation/recombination processes inside of it. (**b**) Generation–recombination via localized trap levels usually described by Shockley–Hall–Read theory

Compact sensors based on the Hall effect are used in a wide variety of products and applications. In addition to magnetometer applications, Hall sensors are also very often used as position sensors (Fig. A.15b).

A.2.7 Continuity Equations

The dynamics or time-dependent behavior of free charge carriers in semiconductors is governed by the basic principle of *conservation of charge* and this leads to a continuity equation that must be obeyed for electrons and holes.

Consider a one-dimensional slice of a semiconductor, dx, as shown in Fig. A.16a: The rate of change of electron concentration in this thin slice will depend only on the net current flowing into it and any generation or recombination (the opposite process) of carriers inside it. Mathematically, the continuity equations for both electrons and holes are given by

$$\frac{\partial n}{\partial t} = \frac{1}{q}\frac{\partial J_n}{\partial x} + (G_n - R_n); \frac{\partial p}{\partial t} = -\frac{1}{q}\frac{\partial J_p}{\partial x} + \left(G_p - R_p\right) \tag{A.46}$$

where G and R represent the generation and recombination rates (per unit volume) for either or electrons or holes. The expressions we have for electron and hole

current densities [Eq. (A.43)] can be substituted into the continuity equations to obtain partial differential equations which can in principle be solved to obtain the temporal and spatial dependence of the carrier concentrations in a semiconductor.

Generation and Recombination

We have seen that excess carriers can be created or generated in a semiconductor by exciting carriers from the valence band to the conduction band. Similarly, excess carriers can recombine by the reverse process. In thermal equilibrium, generation and recombination are in perfect balance. Deviations from thermal equilibrium will lead to either net generation or recombination of carriers in a semiconductor.

For indirect semiconductors such as Ge and Si recombination/generation usually occurs via *localized states* within the band gap caused by impurities and/or defects (*Shockley–Hall–Read* (SHR) recombination[36]) as illustrated in Fig. A.16b. These states act as "stepping stones" for carriers and are referred to as traps or *recombination centers*. On the other hand, for direct band gap semiconductors recombination from the conduction band to valence band without an intermediate state is more probable. The rate of recombination is proportional to the density of carriers, the density of empty available states, and the probability that a carrier passing near an available state will be captured by it.

The net recombination rate, R–G, of carriers in a semiconductor is characterized by the *excess carrier lifetime*, τ, whose value can vary widely depending on the factors listed above and also the physics of the specific recombination mechanism(s) taking place.

We can define the excess electron and hole concentrations in a semiconductor as

$$\begin{aligned} n' &\equiv n - n_0 \\ p' &\equiv p - p_0 \end{aligned} \tag{A.47}$$

where n_0 and p_0 denote the thermal equilibrium carrier concentrations.

If the condition of *low-level injection* is valid (i.e., the majority carrier concentration is unaffected and remains near its thermal equilibrium value), the net recombination rate is simply proportional to the excess minority carrier concentration divided by the minority carrier lifetime.

The continuity equations now become

$$\begin{aligned} \frac{\partial n'}{\partial t} &= \frac{1}{q}\frac{\partial J_n}{\partial x} - \frac{n'}{\tau_n} \\ \frac{\partial p'}{\partial t} &= -\frac{1}{q}\frac{\partial J_p}{\partial x} - \frac{p'}{\tau_p} \end{aligned} \tag{A.48}$$

for the minority carriers in p- and n-type semiconductors, respectively.

[36] R.N. Hall, Phys. Rev. **87**, 387 (1952); W. Shockley, W. T. Read, Jr., Phys. Rev. **87**, 835 (1952).

If low-level injection is not valid the situation becomes somewhat more complicated. For SHR theory the net recombination rate is

$$U = \frac{\sigma_n \sigma_p v_{\text{th}} N_{\text{t}} \left(pn - n_{\text{i}}^2\right)}{\sigma_n \left[n + n_{\text{i}} \exp\left(\frac{E_{\text{t}} - E_{\text{i}}}{k_B T}\right)\right] + \sigma_p \left[p + n_{\text{i}} \exp\left(\frac{E_{\text{i}} - E_{\text{t}}}{k_B T}\right)\right]} \tag{A.49a}$$

where N_{t} is the density of recombination traps and E_{t} their energy, and σ_n and σ_p are the cross sections associated with electron and hole recombination, respectively. U is peaked around $E_{\text{t}} = E_{\text{i}}$, and thus traps near the middle of the band gap tend to make the largest contribution. If only these traps are considered Eq. (A.49a) reduces to

$$U = \frac{\sigma_n \sigma_p v_{\text{th}} N_{\text{t}} \left(pn - n_{\text{i}}^2\right)}{\sigma_n (n + n_{\text{i}}) + \sigma_p (p + n_{\text{i}})} \tag{A.49b}$$

Lastly, a further assumption that is often used to allow approximate calculations is that $\sigma_n = \sigma_p = \sigma$, and thus U simplifies to

$$U = \frac{\sigma v_{\text{th}} N_{\text{t}} \left(pn - n_{\text{i}}^2\right)}{n + p + 2n_{\text{i}}} \tag{A.49c}$$

A.3 Outline of Semiconductor Planar Processing (Optional)

The technology behind virtually all solid-state electronic devices and integrated circuits currently produced is based on planar processing[37] of single-crystal (usually silicon) semiconductor wafers in a sequence of steps, layer by layer, to achieve the desired final structure. The technology of semiconductor electronics is not required for understanding the main concepts in the text and therefore this section is optional reading. However, the ideas and techniques presented will certainly be very helpful for providing underlying insight and putting the various device structures discussed in this book into context. Practically, of course, how electronic devices and integrated circuits are made is very important.

A.3.1 Basic Steps in Processing a Wafer

1. Crystal growth and wafer formation
2. Patterning
3. Dopant deposition

[37] Many of the initial planar device processing concepts were disclosed by Hoerni in 1959.

4. Dielectric formation
5. Etching
6. Metal deposition
7. Formation of interconnect layers
8. Packaging and testing

Note that steps 2–7 are largely applied in a *parallel* fashion across the entire surface of the wafer—it thus costs roughly the same (in time and resources) to manufacture a chip containing billions of components as it does to produce a single discrete device. A brief description of each step follows.

High-purity single-crystal semiconductors are required for precisely controlling electronic device properties, which is necessary for producing reliable large-scale integrated circuits. To obtain a silicon wafer (Fig. A.17a), the starting material is silica or SiO_2 (found in sand). This is reduced in a high-temperature arc furnace containing carbon, resulting in about 90 % pure silicon material. This powdered form of silicon is then reacted with hydrogen chloride to form trichlorosilane liquid. The liquid is then purified via distillation and subsequently reduced with hydrogen to form high-purity (>99.999999 %) polycrystalline silicon. To go from polysilicon to defect-free single crystals the so-called *Czochralski* crystal growth method is standard. Here, a small seed crystal is dipped into a Si melt and slowly withdrawn to grow a progressively larger crystal referred to as in ingot. By incorporating a controlled amount of dopant atoms into the melt doped single crystals can also be grown using the Czochralski method.[38] To create wafers, the single-crystal ingots are typically sliced using a diamond saw followed by polishing and chemical etching to prepare the surface for IC processing.[39] The standard wafer diameter for electronics today is 12 in. (300 mm) with 18 in. (450 mm) soon to be implemented.

Photolithography (Fig. A.17b) is the standard approach used to create the individual patterns on the surface of a wafer that result in the high density of devices inside an integrated circuit. By exposing and developing a photosensitive polymer, known as photoresist, very high-resolution features can be printed on

[38] Silicon grown using the Czochralski method contains significant amounts of oxygen ($\sim 10^{18}$ cm^{-3}) arising from the crucible used to hold the molten Si. The oxygen is usually electrically inactive but can be used to trap unwanted impurities (gettering). For some applications (e.g., some high-power devices) the whole wafer thickness is required and the oxygen atoms can be problematic. This requires a different approach that does not employ a fixed pool of molten material but rather uses a heating element that moves along a polycrystalline rod of material in order to induce crystallization, known as the *Float-zone* technique.

[39] An additional step that involves growing a thin (~micron) layer of crystalline material on top of the underlying single crystal substrate, known as *epitaxy*, is also often performed. This can involve the same type of semiconductor (homoepitaxy, e.g., Si on Si) or different semiconductors (heteroepitaxy, e.g., compound semiconductors or Ge on Si). Vapor phase epitaxy or chemical vapor deposition (CVD) is the most commonly used approach to form the epitaxial or "epi" layers on a silicon wafer. High-quality epi layers allow more precise control of the semiconductor properties and can lead to improved performance of the active devices in an IC.

the surface of a wafer using a predefined mask containing the desired pattern. This type of patterning step is used several times for the various levels of an integrated circuit. The resolution of optical lithography is limited by diffraction, i.e., by the wavelength of radiation used (analogous to an optical microscope). Reducing the wavelength is therefore one way to define finer features and today 193-nm photolithography (deep UV) is standard in industry. Other innovations such as *phase-shift masks* (defining sharper features by modifying the mask to produce destructive interference of light waves for adjacent patterns), *immersion lithography* (reducing wavelength by increasing the refractive index), and *multiple patterning* (superimposing more than one lithography step per level to create a higher

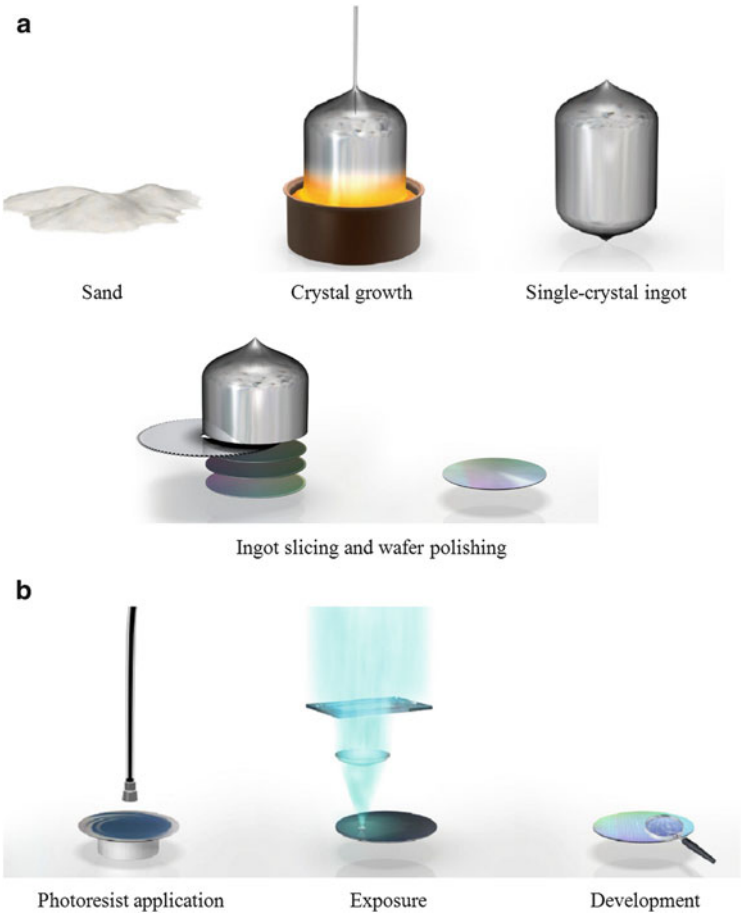


Fig. A.17 Planar processing steps. (Intel Corp.) (**a**) Crystal growth and wafer formation. (**b**) Photolithography. (**c**) Device structures are created via a series of patterning, thin film deposition, and etching steps to create active devices layers, contacts, isolation, etc. (**d**) Metallic interconnect network formation. Cross section of a finished chip with nine metal layers is shown (transistors are small structures at the bottom). (**e**) IC packaging

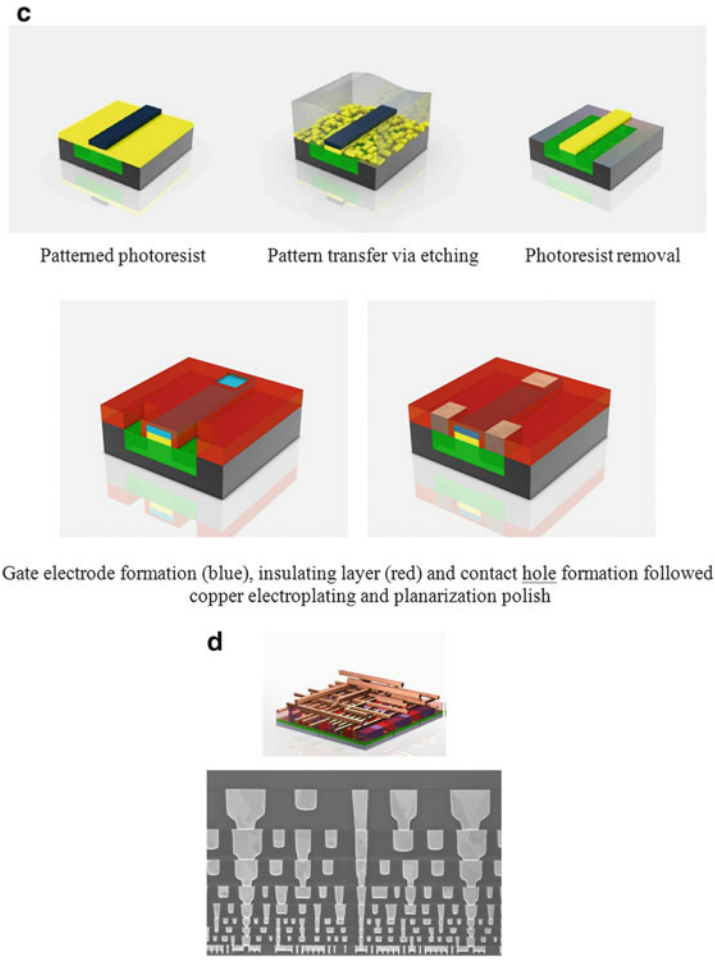


Fig. A.17 (continued)

density of patterns) have allowed state-of-the-art photolithography to reach sub-20-nm feature sizes. Photolithography based on 13.5 nm radiation (Extreme UV) incorporating reflective optics is also being developed and could be used for electronics manufacturing within 5–10 years.

After pattern transfer to the photoresist layer, dopant atoms can be selectively deposited on the exposed Si regions to define the doping levels (and junctions) needed for devices in an IC. The two main methods of depositing dopants are *gas-phase deposition* (or diffusion, for larger/deeper doped regions) and *ion implantation* (accelerated dopant ion species allowing thinner layers with more precise control). Following dopant deposition the remaining photoresist is removed and the resulting high-density patterned regions form the basis of the devices (e.g., transistors) that will make up the IC.

Fig. A.17 (continued)

The next several steps (4, 5, and 6) are combined with photolithography to create the necessary insulating layers (e.g., gate dielectrics (high-κ), device isolation (Si oxidation), etc.), structures (e.g., channels, gates, source–drain), and metallic contacts, respectively, to each device on the wafer surface (see Fig. A.17c for an example process flow). Note that *planarization* steps are performed at various stages of the fabrication process in order to properly prepare and ensure a uniform surface for the next level of processing.

Steps 2–6 are generally known as *front-end processing* of an integrated circuit wafer. Step 7 is the last important IC fabrication step that involves connecting all the individual transistors into the desired circuit functionality and system architecture using metallic wires, which are formed using combinations of patterning/deposition/etching steps. Since ICs have now become very dense with many elements the interconnect superstructure above a wafer can be very complex and usually contains ~10 or more levels on the most advanced chips (Fig. A.17d).

Once all the interconnect layers and intervening insulators have been fabricated, the packaging and test process flow begins. This involves electrically testing each "die" or chip pattern on the wafer, followed by wafer[40] slicing (along the single-crystal planes) and packaging of each individual chip into the final "plug-in" chip

[40] Although standard IC wafers are usually several hundred microns thick, only the top few microns are used to create the devices for an integrated circuit, with the remainder of the wafer acting mainly as a mechanical support.

module (including mechanical protection and thermal dissipation layers) for consumer use (Fig. A.17e). A final sequence of higher level testing is performed prior to product shipment to the end user.

References

1. Ohanian, H.C.: Principles of Quantum Mechanics. Prentice-Hall (1989)
2. Sakurai, J.J: Modern Quantum Mechanics, Revised Edition, Addison-Wesley (1993)
3. Smith, R.A: Semiconductors, 2nd Edn. Cambridge University Press (1978)
4. Kittel, C.: Introduction to Solid State Physics, 8th Edn. Wiley, New Jersey (2005)
5. Kittel, C., Kroemer, H.: Thermal Physics, 2nd Edn. Freeman (1980)
6. Muller, R.S., Kamins, T.I.: Device Electronics for Integrated Circuits, 3rd Edn. Wiley, New York (2003)
7. Ng, K.K: Complete Guide to Semiconductor Devices, 2nd Edn. Wiley Interscience, New York (2002)

Problems

1. *Electron wavelength*. Find the wavelength of free electrons at room temperature.
2. *Semiconductor conductivity*. (1) A silicon sample contains $N_d = 10^{16}\text{cm}^{-3}$ and $N_a = 10^{16}\text{cm}^{-3}$. Calculate the conductivity of the sample. (2) Two semiconductor crystal samples are being examined at room temperature: Sample A has a large concentration of electrons in the conduction band, $n \gg n_i$, while sample B has $n = p = n_i$. Which sample has the highest electron mobility? (3) An n-type Si sample has a resistivity of 0.5 Ω-cm. Is the doping level uniquely defined?
3. *Carrier concentration and drift current*. Find the thermal equilibrium carrier concentrations for a silicon sample doped with $3 \times 10^{17}\text{cm}^{-3}$ phosphorus atoms and $3 \times 10^{17}\text{cm}^{-3}$ arsenic atoms. Sketch the thermal equilibrium band edge diagram and find the position of the Fermi level. What is the maximum drift current density that can flow through this sample?

Appendix B: Useful Data

B.1 General Comments

When solving problems in this textbook you should assume room temperature and that the semiconductor is silicon, unless otherwise stated.

It is important, as always, to use consistent units when solving problems. Electron volts[41] and centimeters are most often used in the electronic devices field; however SI, cgs, or other systems are also perfectly valid and sometimes preferable.

B.2 Fundamental Constants and Definitions

Planck's constant	$h = 6.63 \times 10^{-34}\text{J} - \text{s}; \hbar = \frac{h}{2\pi} = 1.055 \times 10^{-34}\text{J} - \text{s}$
Speed of light	$c = 3 \times 10^{8}$ m/s
Electron/proton charge magnitude	$q = 1.6 \times 10^{-19}\text{C}$
Free electron rest mass	$m_0 = 9.11 \times 10^{-31}\text{kg}$
Proton rest mass	$M_\text{p} = 1.67 \times 10^{-27}\text{kg}$
Avogadro's number	$N_\text{A} = 6.02 \times 10^{23}/\text{mol}$
Atomic mass unit	$\text{amu} = 1.66 \times 10^{-27}\text{kg}$
Boltzmann constant	$k_\text{B} = 1.38 \times 10^{-23}\text{J/K}$
Permittivity of free space	$\varepsilon_0 = 8.85 \times 10^{-12}\text{F/m}$

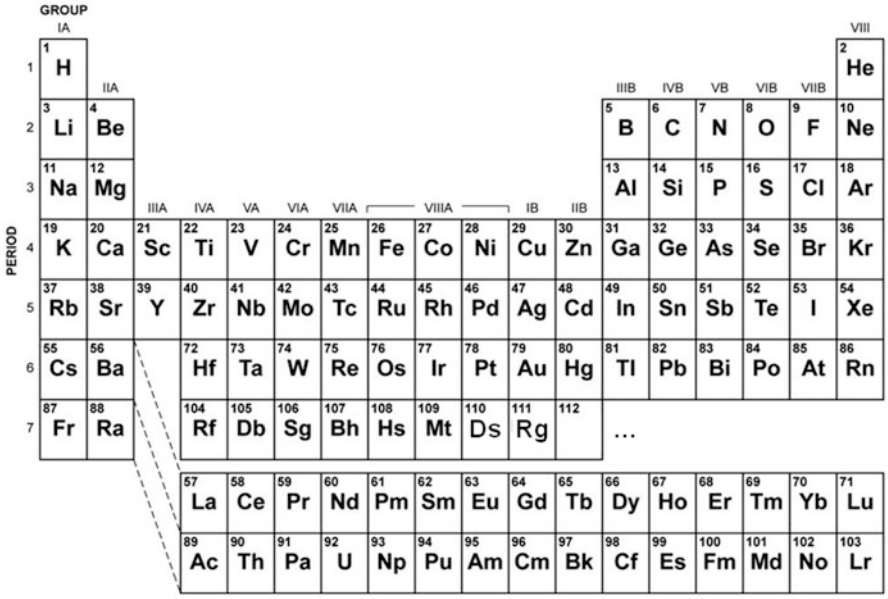
$$1\ \text{eV} = 1.6 \times 10^{-19}\ \text{J}; \quad k_\text{B}T(300\ \text{K}) \approx 0.026\ \text{eV}; \qquad 1\ \text{Å} = 10^{-10}\text{m} = 0.1\ \text{nm}$$

Photon energy: $E\ (\text{eV}) = 1.24/\lambda\ (\mu\text{m})$

[41] Recall that 1 eV is the amount of energy gained by an electron that is accelerated through a potential difference of 1 V (or 1 J/C). Therefore, the term qV that often appears in electronic device work gives an energy in eV if the bias is in volts.

C. Papadopoulos, *Solid-State Electronic Devices: An Introduction*,
Undergraduate Lecture Notes in Physics, DOI 10.1007/978-1-4614-8836-1,

B.3 Periodic Table



B.4 Semiconductors and Other Relevant Electronic Materials

Important parameters[42]

[42] All parameters at 300 K.

Crystal	E_g (eV)[a]	n_i (cm^{-3})	Effective mass[b] (in units of m_0)	Intrinsic mobility ($cm^2/V\,s$)	Effective density of states (cm^{-3})	Relative permittivity ε_r	Electron affinity qX (eV)	Lattice constant[c] a (nm)
Diamond	i	–	$m_n = 0.52$	$\mu_n = 2100$	$N_c = 9 \times 10^{18}$	5.7	–	0.3567
	5.5		$m_n = 0.40$	$\mu_p = 2000$	$N_v = 1.6 \times 10^{19}$			
			$m_p = 0.75$					
			$m_p = 0.46$					
Si	i	1.02×10^{10}#	$m_n = 1.08$	$\mu_n = 1450$	$N_c = 2.8 \times 10^{19}$	11.7	4.05	0.5431
	1.12		$m_n = 0.26$	$\mu_p = 450$	$N_v = 1.0 \times 10^{19}$			
			$m_p = 0.81$					
			$m_p = 0.386$					
Ge	i	2.33×10^{13}	$m_n = 0.55$	$\mu_n = 3900$	$N_c = 1.0 \times 10^{19}$	16.0	4.0	0.5658
	0.67		$m_n = 0.12$	$\mu_p = 1900$	$N_v = 6.0 \times 10^{18}$			
			$m_p = 0.3$					
			$m_p = 0.21$					
GaAs	d	2.1×10^{6}	$m_n = 0.067$	$\mu_n = 8500$	$N_c = 4.7 \times 10^{17}$	13.1	4.07	0.5653
	1.42		$m_p = 0.47$	$\mu_p = 400$	$N_v = 7.0 \times 10^{18}$			
			$m_p = 0.34$					
GaP	i	–	$m_n = 0.79$	$\mu_n = 250$	$N_c = 1.9 \times 10^{19}$	11.1	3.8	0.5450
	2.24		$m_n = 0.35$	$\mu_p = 150$	$N_v = 1.2 \times 10^{19}$			
			$m_p = 0.83$					
			$m_p = 0.50$					
GaN	d	–	$m_n = 0.2$	$\mu_n = 900$	$N_c = 2.2 \times 10^{18}$	9.7	4.1	0.3190
	3.4		$m_p = 1.5$	$\mu_p = 200$	$N_v = 4.6 \times 10^{19}$			0.5189
			$m_p = 0.8$					
InP	d	1.3×10^{7}	$m_n = 0.073$	$\mu_n = 4600$	$N_c = 5.2 \times 10^{17}$	12.4	4.35	0.5869
	1.34		$m_p = 0.6$	$\mu_p = 150$	$N_v = 1.1 \times 10^{19}$			
InSb	d	2×10^{16}	$m_n = 0.012$	$\mu_n = 77000$	$N_c = 4.2 \times 10^{16}$	17.9	4.6	0.6479
	0.17		$m_p = 0.43$	$\mu_p = 850$	$N_v = 7.3 \times 10^{18}$			

(continued)

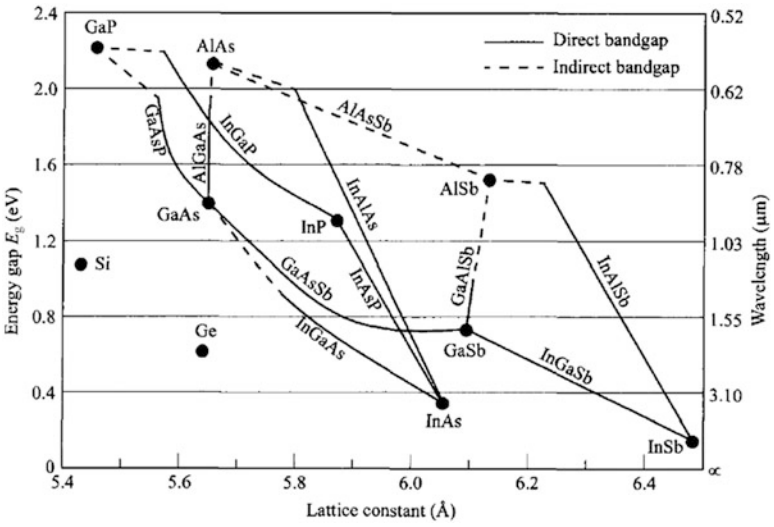
Crystal	E_g (eV)[a]	n_i (cm^{-3})	Effective mass[b] (in units of m_0)	Intrinsic mobility (cm^2/V s)	Effective density of states (cm^{-3})	Relative permittivity ε_r	Electron affinity qX (eV)	Lattice constant[c] a (nm)
InAs	d 0.36	1×10^{15}	$m_n = 0.023$ $m_p = 0.41$	$\mu_n = 33000$ $\mu_p = 450$	$N_c = 8.3 \times 10^{16}$ $N_v = 6.4 \times 10^{18}$	14.6	4.9	0.6058
Graphene	0	8×10^{10} (cm^{-2})	–	$\sim 10^5$	–	2.2	4.6	0.246
CNT	d 0 ~ 2	0 ~ 10^7 (cm^{-1})	–	$\sim 10^5$	–	–	~4.5	–
SiO_2	8–9	–	–	–	–	3.9	0.95	–
HfO_2	5.8	–	–	–	–	20–25	2.1	0.508

[a]d—direct band gap; i—indirect band gap.
[b]Where two values are listed for the same parameter, the first value should be used for density of states calculations and the second for conductivity calculations.
[c]Two lattice constants (a, c) correspond to materials that crystallize in the hexagonal wurtzite structure.
[#]A value of ~1.5×10^{10} cm^{-3} is often used instead for calculations/problems involving Si.

Metal work functions

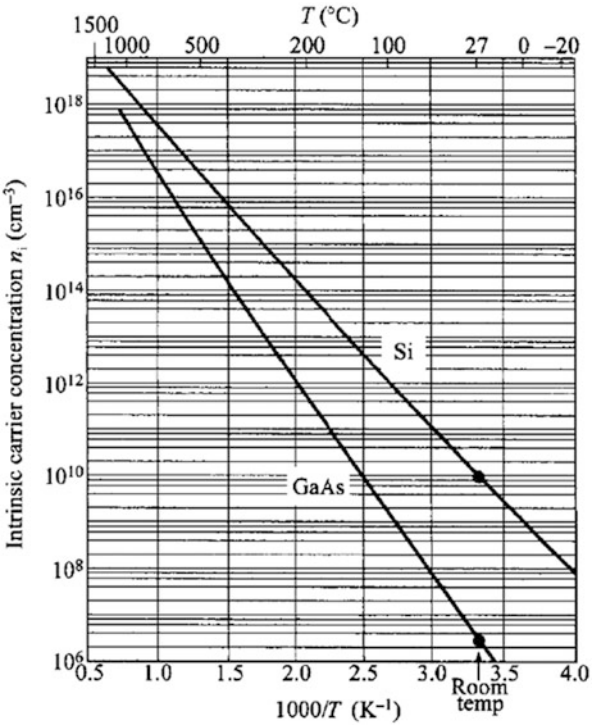
Metal	Work function $q\Phi_M$ (eV)
Al	4.1
Au	5.1
Ag	4.75
Cu	4.7
Pt	5.3
Ti	4.3
Ta	4.25
W	4.5
Ni	5.05
Fe	4.7
Co	5.0

Band gap energy vs. lattice constant [4]

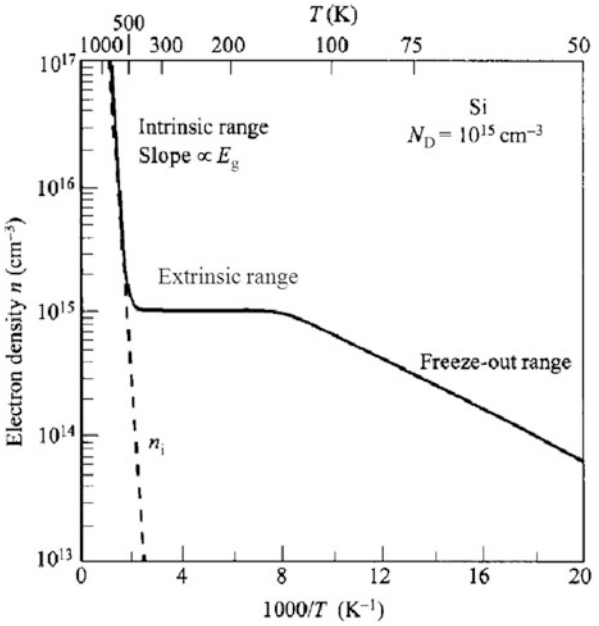


B.5 Silicon Properties

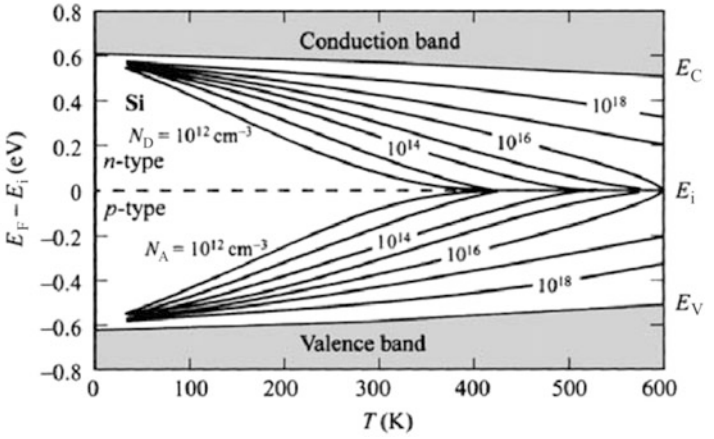
Intrinsic carrier concentration vs. temperature [4]



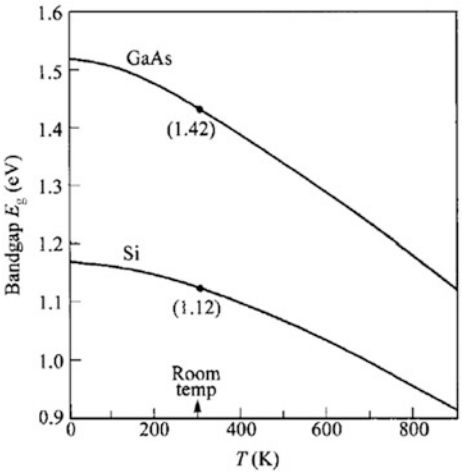
Electron concentration vs. *temperature* (After R. A. Smith, *Semiconductors*, 2nd Edition, Cambridge University Press, 1978.)



Fermi level vs. *temperature* (After [5])



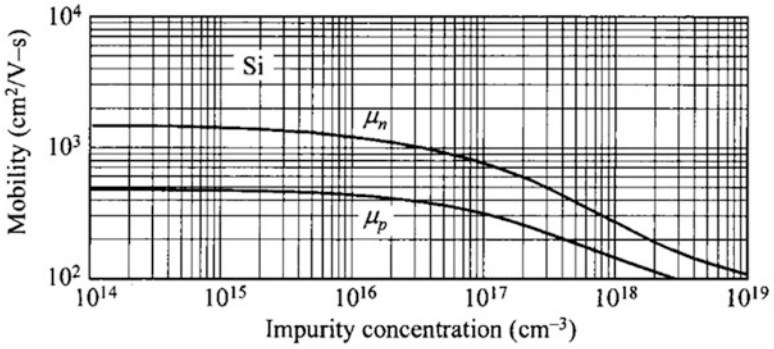
Note that the band gap decreases with temperature in the above diagram. This is a typical trend in most semiconductors as the average lattice spacing increases with temperature [4]:



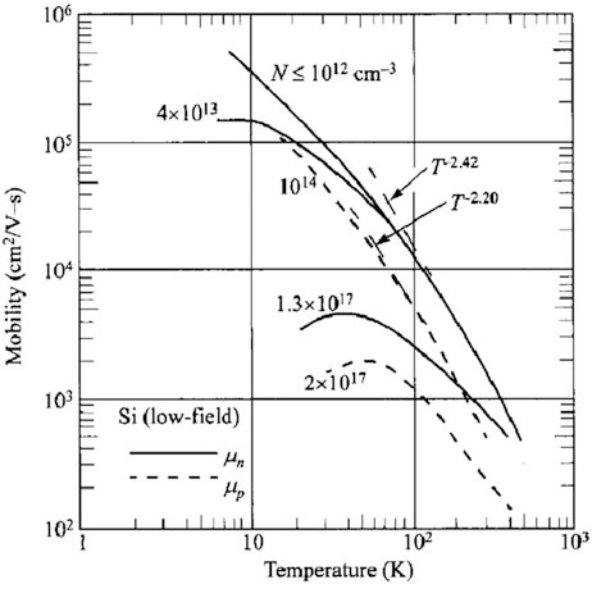
	$E_g(0)$ (eV)	α (eV/K)	β (K)
GaAs	1.519	5.4×10^{-4}	204
Si	1.169	4.9×10^{-4}	655

$$E_g(T) = E_g(0) - \frac{\alpha T^2}{T+\beta}$$

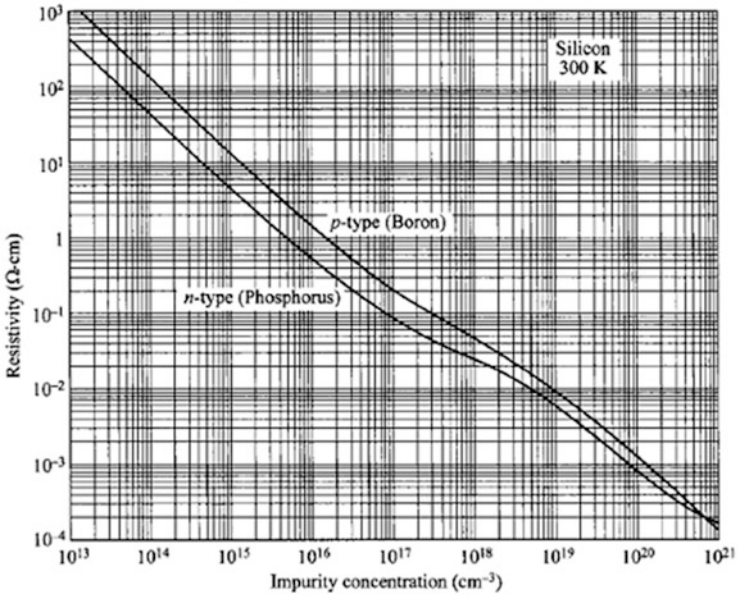
Carrier mobility vs. *total impurity concentration* (300 K) [4]

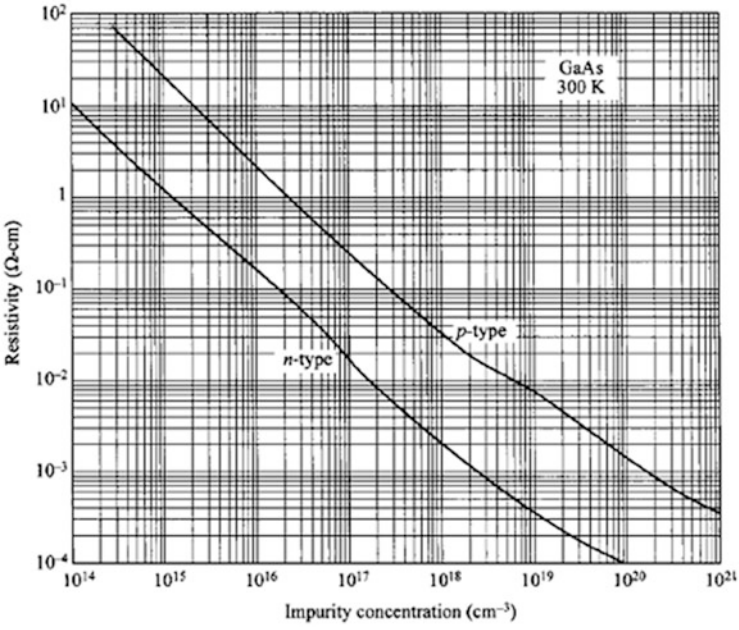


Carrier mobility (low-field) in Si as a function of temperature (After C. Jacobini, C. Canali, G. Ottaviani, A. A. Quaranta, Solid-State Electron. **20**, 77 (1977).)

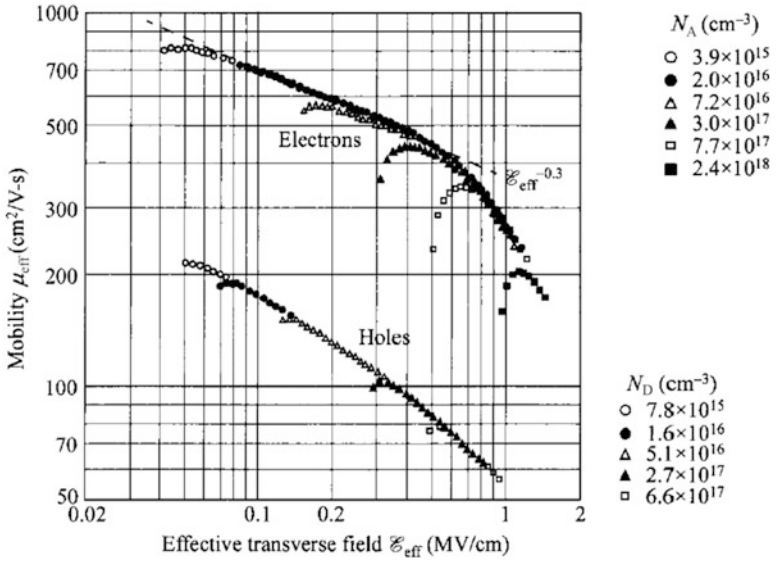


Resistivity vs. *dopant concentration* (only one type of dopant is present for each curve, i.e., the data are for uncompensated material) (After [4])

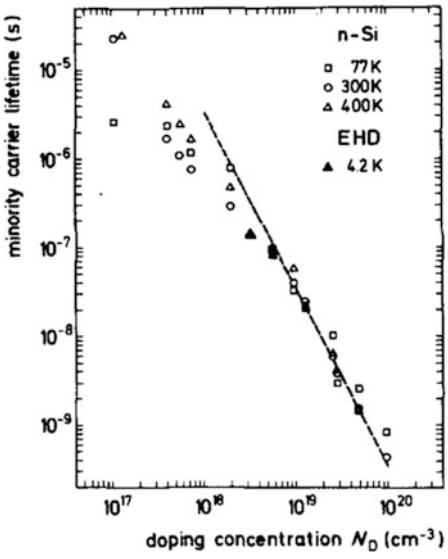


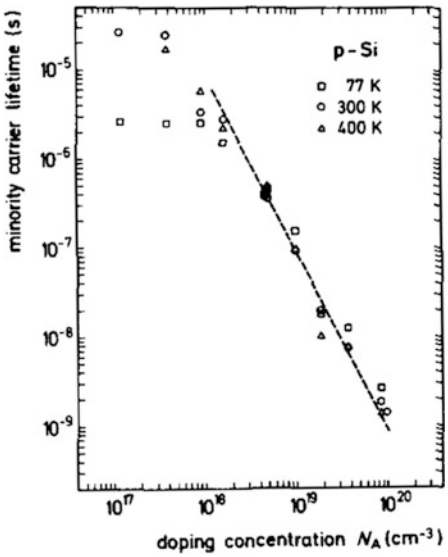


MOS inversion layer carrier mobilities for silicon (100) surface vs. *transverse electric field* [4]

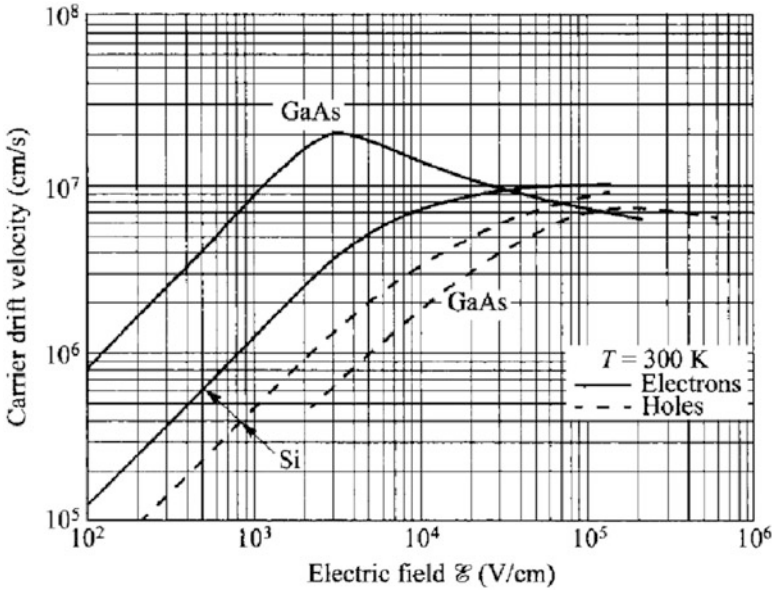


Minority carrier lifetimes versus *doping level for n- and p-type silicon* (After J. Dziewior, W. Schmid, Appl. Phys. Lett. **31**, 346 (1977).)





Carrier drift velocity vs. *applied electric field* [4]



References

1. Kittel, C.: Introduction to Solid State Physics, 7th Edn. Wiley, New Jersey (1996)
2. Martienssen W., Warlimont, H. (eds.): Springer Handbook of Condensed Matter and Materials Data. Springer, Berlin (2005)
3. Muller, R.S., Kamins, T. I.: Device Electronics for Integrated Circuits, 3rd Edn. Wiley, New York (2003)
4. Sze, S.M., Ng, K.K.: Physics of Semiconductor Devices, 3rd Edn. Wiley Interscience, New Jersey (2007)
5. Grove, A.S.: Physics and Technology of Semiconductor Devices. Wiley (1967)

List of Symbols

a	Lattice constant
$\mathring{A}$	Angström
a_0	Bohr radius
B	Magnetic field
BV_{CB0}	Collector-base breakdown voltage with emitter open-circuited
BV_{CE0}	Collector-emitter breakdown voltage with base open-circuited
C	Small-signal capacitance per unit area
c	Speed of light
C_d	Diffusion capacitance per unit area
C_j	Junction capacitance per unit area
C_{min}	Minimum MOS capacitance
C_{ox}	Oxide capacitance per unit area
C_s	MOS depletion capacitance
$D(E)$	Electronic density of states
D_n	Electron diffusion coefficient
D_p	Hole diffusion coefficient
E_0	Vacuum energy level
E_c	Conduction band edge energy
E_{cn}	Conduction band edge energy in n-type semiconductor
E_{cp}	Conduction band edge energy in p-type semiconductor
E_F	Fermi level
E_g	Band gap energy
E_H	Hall field
E_i	Intrinsic Fermi level
E_{max}	Maximum electric field
E_n	Energy eigenvalues
E_{ox}	Oxide electric field
E_{s0}	Electric field at oxide-silicon interface
E_t	Recombination trap energy level
E_v	Valence band edge energy
E_x	Electric field in the x-direction
$f(E)$	Fermi-Dirac distribution

(continued)

C. Papadopoulos, *Solid-State Electronic Devices: An Introduction*,
Undergraduate Lecture Notes in Physics, DOI 10.1007/978-1-4614-8836-1,

(continued)

FF	Fill factor
f_T	Cut-off frequency
G	Small-signal conductance; carrier generation rate
g_0	Output transistor conductance
g_d	Output drain conductance
g_m	Transconductance
$g_{m\mathrm{sat}}$	Saturation transconductance
GN	Gummel number
H	Energy operator (Hamiltonian)
h	Planck's constant
$\hbar$	$\frac{h}{2\pi}$
I	Current
I'_0	Effective reverse saturation current
I_0	Reverse saturation current
I_B	Base current
I_C	Collector current
I_CB0	Collector current with emitter open-circuited
I_CE0	Collector current with base open-circuited
I_D	Drain current
I_D0	Subthreshold drain saturation current
I_dark	Diode current in the absence of light
I_Dsat	MOSFET saturation current
I_E	Emitter current
I_F	Forward current flowing through a diode
I_F0	Ebers-Moll forward current parameter
I_GR0	Reverse saturation current due to space-charge generation-recombination
I_L	Diode photocurrent
I_R	Reverse current flowing through a diode
I_R0	Ebers-Moll reverse current parameter
I_rB	Base recombination current
I_sc	Diode short-circuit current
J	Current density
J_0	Reverse saturation current density
J_G	Space-charge generation current density
J_ideal	Ideal diode current density
J_1	Leakage current density
J_n	Electron current density
J_p	Hole current density
J_R	Space-charge recombination current density
J_x	Current density in the x-direction
k	Wave vector; spring constant
k_B	Boltzmann constant
L	Channel length
L_D	Debye length
L_n	Electron diffusion length
L_p	Hole diffusion length
M	Multiplication factor for avalanching

(continued)

(continued)

m	Mass
m^*	Effective mass
m_0	Free electron mass
m_n	Electron effective mass
m_p	Hole effective mass
n	Electron carrier concentration
n'	Excess electron concentration
n'_{p}	Excess electron concentration in p-type semiconductor
n_0	Thermal equilibrium electron concentration
N_{a}	Acceptor impurity concentration
N_{c}	Effective density of states in the conduction band
N_{d}	Donor impurity concentration
N_{I}	Implanted dopant ions per unit area
n_{i}	Intrinsic carrier concentration
n_{n}	Electron concentration in n-type semiconductor
n_{n0}	Thermal equilibrium electron concentration in n-type semiconductor
n_{p}	Electron concentration in p-type semiconductor
n_{p0}	Thermal equilibrium electron concentration in p-type semiconductor
n_{s}	Electron density in semiconductor near an interface
N_{t}	Recombination trap density
N_{v}	Effective density of states in the valence band
p	Hole carrier concentration
P	Power
p'	Excess hole concentration
p'_{n}	Excess hole concentration in n-type semiconductor
p_0	Thermal equilibrium hole concentration
P_{in}	Input power
P_{m}	Maximum output power
p_{n}	Hole concentration in n-type semiconductor
p_{n0}	Thermal equilibrium hole concentration in n-type semiconductor
p_{p}	Hole concentration in p-type semiconductor
p_{p0}	Thermal equilibrium hole concentration in p-type semiconductor
p_{s}	Hole density in semiconductor near an interface
q	Electron/proton charge magnitude
Q_{B}	Majority carrier charge density per unit area in the base of a bipolar transistor
Q_{d}	Depletion space-charge density per unit area
Q_{dmax}	Maximum depletion space-charge density per unit area
Q_{f}	Fixed oxide interface charge density per unit area
Q_{fg}	Floating-gate charge per unit area
Q_n	Excess minority electron charge per unit area; MOS inversion layer charge per unit area
Q_p	Excess minority hole charge per unit area; MOS inversion layer charge per unit area
Q_{s}	Space charge per unit area
R	Resistance; carrier recombination rate
R_{H}	Hall coefficient
R_{s}	Series resistance
T	Temperature; tunneling probability
t_{r}	Diode recovery time

(continued)

(continued)

t_s	Diode storage time
T_{tr}	Channel transit time
U	Net recombination rate
V	Voltage
$V(x)$	Potential energy
$\overline{v}$	Drift velocity
V_A	Early voltage
V_a	Applied voltage
V_B	Substrate bias
V_{BC}	Base-collector voltage of a bipolar transistor
V_{BE}	Base-emitter voltage of a bipolar transistor
V_{bi}	*pn* junction built-in potential
V_{BR}	Breakdown voltage
V_C	Channel bias
V_{CEsat}	Voltage drop from collector to emitter in saturation
V_D	Drain voltage
V_{Dsat}	MOSFET saturation voltage
V_{FB}	Flat-band voltage
V_G	Gate voltage
V_H	Hall voltage
V_n	Voltage at edge of depletion region in n-type semiconductor
V_{oc}	Diode open-circuit voltage
V_{ox}	Voltage drop across the oxide of a MOS structure
V_p	Voltage at edge of depletion region in p-type semiconductor
V_S	Source voltage
v_{sat}	Saturation velocity
V_T	MOS threshold voltage
V_t	Thermal voltage
v_{th}	Thermal velocity
W	Channel width
W_B	Distance from metallurgical junction to contact on n-side of *pn* junction
W_E	Distance from metallurgical junction to contact on p-side of *pn* junction
x_B	Neutral width of base region; n-region of a *pn* junction
x_d	Width of depletion region
x_{dmax}	Maximum width of the depletion region
x_E	Neutral width of emitter region; p-region of a *pn* junction
x_n	Extent of depletion region in n-type semiconductor
x_{ox}	Oxide thickness
x_p	Extent of depletion region in p-type semiconductor
α	Bulk-charge factor
α_F	Forward-active dc alpha
α_T	Base transport factor
β	Bipolar current gain
β_F	Forward-active current gain
β_R	Reverse-active current gain
γ	Emitter injection efficiency
δ	Fraction of charge removed during diode storage time

(continued)

(continued)

ε_0	Vacuum permittivity
ε_{ox}	Oxide permittivity
ε_r	Relative permittivity or dielectric constant
ε_s	Semiconductor permittivity
η	Diode ideality factor; solar cell power conversion efficiency
λ	Carrier mean free path; wavelength
μ_n	Electron mobility
μ_p	Hole mobility
ν	Frequency
ρ	Electrical resistivity
$\rho(x)$	Distributed charge density
σ	Electrical conductivity
σ_n	Electron recombination cross-section
σ_p	Hole recombination cross-section
τ	Mean free time; excess carrier lifetime
τ_B	Base transit time
τ_C	Collector transit time
τ_{EC}	Bipolar transistor delay time
τ_n	Excess electron lifetime
τ_p	Excess hole lifetime
τ_{RC}	*RC* time constant
τ_t	Transit time
Φ	Work function (voltage)
$\phi(x)$	Potential inside silicon substrate of a MOS structure
ϕ_0	Fermi level pinning (voltage)
ϕ_B	Schottky barrier height (voltage)
ϕ_{bi}	Metal-semiconductor junction built-in potential
Φ_M	Metal work function (voltage)
Φ_{MS}	Metal-semiconductor work function difference, $\Phi_M - \Phi_S$
ϕ_n	Potential in bulk n-type substrate of a MOS structure
ϕ_p	Potential in bulk p-type substrate of a MOS structure
Φ_S	Semiconductor work function (voltage)
ϕ_s	Potential at oxide-silicon interface
X	Electron affinity (voltage)
$\psi(x)$	Spatial part of the wavefunction
$\Psi(x,t)$	Wavefunction
ω	Angular frequency
ω_0	Resonance frequency

General Bibliography

1. Grove, A.S.: Physics and Technology of Semiconductor Devices. Wiley (1967)—A classic in the field.
2. International Technology Roadmap for Semiconductors, ITRS, (2011–2012)—An immense practical resource for the present and future of the semiconductor industry that is released every two years with brief updates in the intervening years.
3. Kittel, C.: Introduction to Solid State Physics, 7th Edn. Wiley, New Jersey (1996)—A standard reference for solid state physics. Now in its 8th edition, earlier versions are also worth a look.
4. Muller, R.S., Kamins, T.I.: Device Electronics for Integrated Circuits, 3rd Edn. Wiley, New York (2003)—Containing practical and fundamental concepts, this textbook emphasizes both an intuitive and comprehensive understanding of the subject.
5. Ohanian, H.C.: Principles of Quantum Mechanics. Prentice-Hall (1989)—A good introduction to fundamental quantum mechanical concepts and their mathematical formalism.
6. Pierret, R.F., Neudeck, G.W.: Modular Series on Solid State Devices Volumes I-IV. Prentice-Hall (1990)—A standard set of introductory level volumes on semiconductor properties and devices.
7. Purcell, E.M.: Electricity and Magnetism, 2nd Edn. McGraw-Hill (1985)—An excellent introduction to classical electromagnetic theory. Chapters 1–4 are particularly relevant to electronic devices. (An updated edition has also recently been published.)
8. Shur, M.S.: Introduction to Electronic Devices. Wiley (1995)—Contains many useful insights by an author who has a clear mastery of the material.
9. Streetman, B., Banerjee, S.: Solid State Electronic Devices, 6th Edn. Prentice-Hall (2005)—A standard beginner- to intermediate-level devices text.
10. Sze, S.M., Ng, K.K.: Physics of Semiconductor Devices, 3rd Edn. Wiley Interscience, New Jersey (2007)—A comprehensive reference text.

C. Papadopoulos, *Solid-State Electronic Devices: An Introduction*,
Undergraduate Lecture Notes in Physics, DOI 10.1007/978-1-4614-8836-1,

Index

C. Papadopoulos, *Solid-State Electronic Devices: An Introduction*, Undergraduate Lecture Notes in Physics, DOI 10.1007/978-1-4614-8836-1,