CAMBRIDGE

Cambridge University Press
978-1-108-48002-4 — Fundamentals of Modern VLSI Devices
Yuan Taur , Tak H. Ning
Excerpt
More Information

1 Introduction

1.1 Evolution of VLSI Device Technology 1
1.2 Scope and Brief Description of the Book 5

Since the invention of the bipolar transistor in 1947, there has been an unprecedented growth of the semiconductor industry, with an enormous impact on the way people work and live. In the last forty years or so, by far the strongest growth area of the semiconductor industry has been in silicon very-large-scale-integration (VLSI) technology. The sustained growth in VLSI technology is fueled by the continued shrinking of transistors to ever smaller dimensions. The benefits of miniaturization – higher packing densities, higher circuit speeds, and lower power dissipation – have been key in the evolutionary progress leading to today's computers, wireless units, and communication systems that offer superior performance, dramatically reduced cost per function, and much reduced physical size, in comparison with their predecessors. On the economic side, the integrated-circuit (IC) business has grown worldwide in sales from $1 billion in 1970 to $20 billion in 1984 and has reached $439 billion in 2020. The electronics industry is now among the largest industries in terms of output as well as employment in many nations. The importance of microelectronics in economic, social, and even political development throughout the world will no doubt continue to ascend. The large worldwide investment in VLSI technology constitutes a formidable driving force that will all but guarantee the continued progress in IC integration density and speed, for as long as physical principles will allow.

1.1 Evolution of VLSI Device Technology

1.1.1 Historical Perspective

An excellent account of the evolution of the metal–oxide–semiconductor field-effect transistor (MOSFET), from its initial conception to VLSI applications in the mid-1980s, can be found in a paper by Sah (1988). Figure 1.1 gives a chronology of the major milestone events in the development of VLSI technology. The vertical bipolar transistor technology was developed early on and was applied to the first integrated-circuit memory in mainframe computers in the 1960s. Vertical bipolar transistors have been used all along where raw circuit speed is most important, for bipolar circuits remain the fastest at the individual-circuit level. However, the large power dissipation

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE
Cambridge University Press
978-1-108-48002-4 — Fundamentals of Modern VLSI Devices
Yuan Taur , Tak H. Ning
Excerpt
More Information

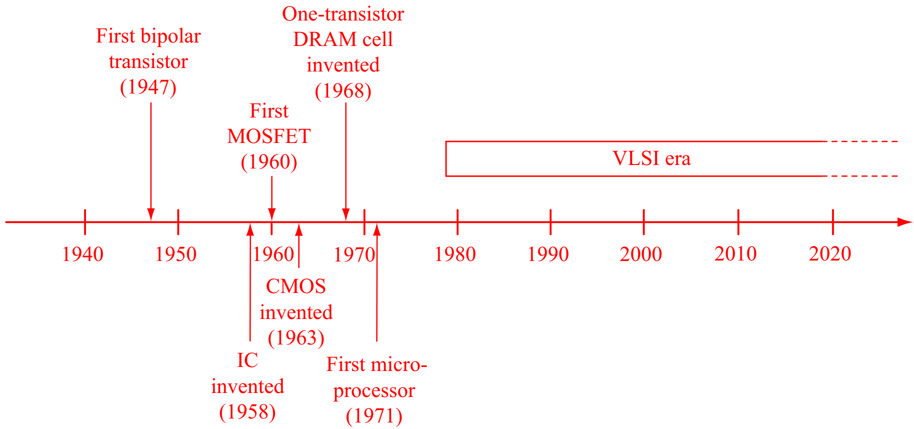

Figure 1.1 A brief chronology of the major milestones in the development of VLSI

of vertical bipolar circuits has severely limited their integration level, to about 10^4 circuits[1] per chip. This integration level is very low by today's VLSI standard.

The idea of modulating the surface conductance of a semiconductor by the application of an electric field was first envisioned in 1930. However, early attempts to fabricate a surface-field-controlled device were not successful because of the presence of large densities of surface states which effectively shielded the surface potential from the influence of an external field. The first MOSFET on a silicon substrate using SiO_2 as the gate insulator was fabricated in 1960 (Kahng and Atalla, 1960). During the 1960s and 1970s, n-channel and p-channel MOSFETs were widely used, along with bipolar transistors, for implementing circuit functions on a silicon chip. Although the MOSFET devices were slow compared to the bipolar devices, they had a higher layout density and were relatively simple to fabricate; the simplest MOSFET chip could be made using only four masks and a single doping step. However, just like vertical bipolar circuits, single-polarity MOSFET circuits suffered from large standby power dissipation, hence were limited in the level of integration on a chip.

The major breakthrough in the level of integration came in 1963 with the invention of CMOS (complementary MOS) (Wanlass and Sah, 1963), in which n-channel and p-channel MOSFETs are constructed side by side on the same substrate. A CMOS circuit typically consists of an n-channel MOSFET and a p-channel MOSFET connected in series between the power-supply terminals, so that there is negligible standby power dissipation. Significant power is dissipated only during switching of the circuit (i.e., only when the circuits are active). By cleverly designing the "switch activities" of the circuits on a chip to minimize active power dissipation, engineers have been able to integrate billions of CMOS transistors on a single chip and still have the chip readily air-coolable. Until the minimum feature size of lithography reached 180 nm, the integration level of CMOS was not limited by chip-level power

[1] ECL circuits, discussed in Section 11.2.

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-1-108-48002-4 — Fundamentals of Modern VLSI Devices
Yuan Taur , Tak H. Ning
Excerpt
More Information

dissipation, but by chip fabrication technology. Another advantage of CMOS circuits comes from the ratioless, full rail-to-rail logic swing, which improves the noise margin and makes a CMOS chip easier to design.

As CMOS scaling reached the 0.5-μm level in the early 1990s, the performance of high-end computers built using CMOS started to approach those built using bipolar, due to the much higher integration level of CMOS chips. Designers of high-end computer systems were able to meet their performance targets using CMOS instead of bipolar (Rao *et al.*, 1997). Since then, CMOS has become the technology for digital circuits, and vertical bipolar is used primarily in radio-frequency (RF) and analog circuits only.

Advances in lithography and etching technologies have enabled the industry to scale down transistors in physical dimensions, and to pack more transistors in the same chip area. Such progress, combined with a steady growth in chip size, resulted in an exponential growth in the number of transistors and memory bits per chip. The technology trends up to 2020 in these areas are illustrated in Figure 1.2. Traditionally, dynamic random-access memories (DRAMs) have contained the highest component count of any IC chips. This has been so because of the small size of the one-transistor memory cell (Dennard, 1968) and because of the large and often insatiable demand for more memory in computing systems. It is interesting to note that the entire content of this book can be stored in one 64-Mb DRAM chip, which was in volume production in 1997 and has an area equivalent to a square of about 1.2×1.2 cm^2.

One remarkable feature of silicon devices that fueled the rapid growth of the information technology industry is that their speed increases and their cost decreases as their size is reduced. The transistors manufactured in 2020 were 10-times faster and occupy less than 1% of the area of those built 20 years earlier. This is illustrated in the trend of microprocessor units (MPUs) in Figure 1.2. The increase in the clock frequency of microprocessors is the result of a combination of improvements in microprocessor architecture and improvements in transistor speed.

1.1.2 Recent Developments

Since the publication of the second edition of this book in 2009, there have been major developments in the VLSI industry. Several fabrication technologies emerging at the time have taken hold, enabling the continued chip-level density improvements, resulting in continued reduction of cost per transistor and cost per memory bit. These in turn have driven the continued growth of the semiconductor industry. These recent developments include the following.

- Immersion lithography has been adopted for volume IC manufacturing (Lin, 2004). Immersion lithography is a photolithography resolution enhancement technique where the usual gap between the final lens and the wafer surface is replaced with a liquid medium having a refractive index greater than one. The resolution enhancement is equal to the refractive index of the liquid used. With immersion, deep ultraviolet (DUV) lithography systems remain the work horse for semiconductor manufacturing today.

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-1-108-48002-4 — Fundamentals of Modern VLSI Devices
Yuan Taur , Tak H. Ning
Excerpt
More Information

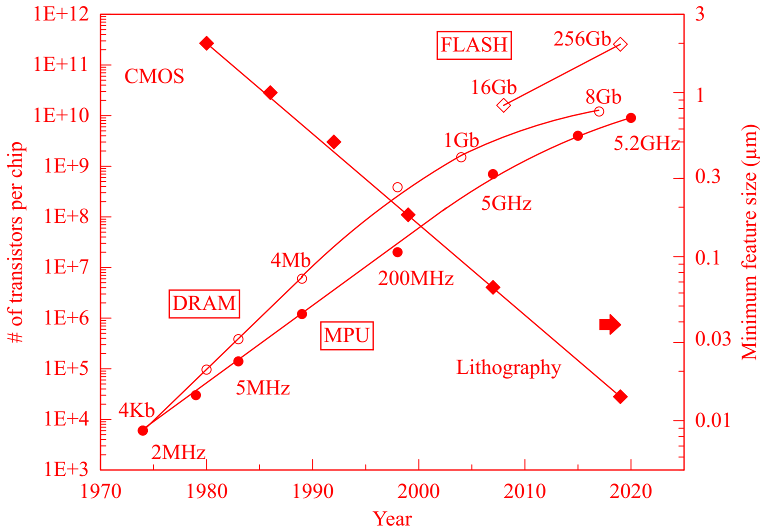

Figure 1.2 Trends in lithographic feature size, number of transistors per chip for DRAM and MPU, and number of memory bits per chip for Flash. The transistor count for DRAM is computed as 1.5-times the number of bits on the chip to account for the peripheral circuits. Data points represent announced leading-edge products

- Driven by the need for low-power and light-weight data storage in battery-operated personal systems, NAND flash (the highest density version of the electrically programmable and erasable nonvolatile memory) development has been on an exceptionally steep trajectory since the mid-1990s. In the past decade or so, NAND flash has overtaken DRAM as the IC chip with the highest component count, as shown in Figure 1.2 (Kim, 2008). Since then, the combination of 3D NAND process technology, where upwards of 100+ layers of NAND flash memory cells are stacked on top of another to form a three-dimensional IC chip, and multi-bit-per-cell design has dramatically increased the chip-level bit density of NAND flash.
- For a long time, a common practice in designing a scaled CMOS device was to allow its off current to increase, by reducing the device threshold voltage, in order to increase the on current to achieve the targeted performance of the scaled device. As a result, the off current of scaled CMOS devices had been increasing from one generation to the next. By the time scaling reached the 65 nm node, the off current of scaled CMOS device had reached 100 nA/µm, the maximum level acceptable to designers of high-end microprocessors. Since then, high-performance CMOS devices have been designed with a nominal off current of 100 nA/µm (Kuhn *et al.*, 2012). Capping the tradeoff between on current and off current in designing scaled CMOS devices severely limits the speed (clock frequency) of microprocessors. Today, the highest speed microprocessors run at 5.2 GHz (Berry *et al.*, 2020), practically the same as those in 2009 (see Figure 1.2).

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-1-108-48002-4 — Fundamentals of Modern VLSI Devices
Yuan Taur , Tak H. Ning
Excerpt
More Information

- Without the ability to increase the device off current, CMOS designers turned to device structures having fully depleted device body, which enables the subthreshold swing of the device to reach the ideal 60 mV/decade at room temperature. Today, FinFET CMOS, where the depleted device body is shaped like a fin, planar ETSOI (extremely thin silicon-on-insulator) CMOS, which is basically fully depleted SOI CMOS, as well as traditional planar bulk CMOS are in volume manufacturing.
- Partially depleted (PD) planar SOI CMOS ran its course over a span of about fifteen years, with the first PD SOI CMOS microprocessor fabricated in the 220 nm node in 1999 (Shahidi *et al*., 1999), and the last in the 22 nm node in 2015 (Freeman *et al*., 2015). PD SOI CMOS is judged not scalable to smaller dimensions. However, SOI wafers suitable for PD SOI CMOS have found new applications in complementary (integration of n−p−n and p−n−p) lateral bipolar (Cai *et al*., 2011), offering interesting possibilities in bipolar for VLSI.

1.2 Scope and Brief Description of the Book

In writing this book, it is our goal to address the factors governing the performance of modern VLSI devices in depth. This is carried out by first discussing the device physics that goes into the design of individual device parameters, and then discussing the effects of these parameters on the performance of small-dimension modern transistors at the basic circuit level. A substantial part of the book is devoted to in-depth discussions on the interdependency among the device parameters and the subtle tradeoffs in the design of modern CMOS and bipolar transistors.

This book contains sufficient background tutorials to be used as a textbook for students taking a graduate or advanced undergraduate course in microelectronics. The prerequisite is one semester of either solid-state physics or semiconductor physics. For the practicing engineer, this book provides an extensive source of reference material that covers the fundamentals of CMOS and bipolar technologies, devices, and circuits. It should be useful to VLSI process engineers and circuit designers interested in learning basic device principles, and to device design or characterization engineers who desire more in-depth knowledge in their specialized areas.

New topics and materials in the third edition include an expanded chapter on ETSOI and FinFETs, non-GCA (Gradual Channel Approximation) model for MOSFETs, and the relatively recent development of symmetric lateral bipolar transistors on SOI. Also added are sections on high-κ gate dielectrics, metal gates, strain effect on mobility, and interface–state models. Much of the materials in the second edition have been restructured by consolidating all the appendices into the main chapters for a more focused coverage of the various subjects. Here is a brief description of each chapter.

Chapter 2: Basic Device Physics

Chapter 2 covers the appropriate level of basic device physics to make the book self-contained, and to prepare the reader with the necessary background on device operation and material physics to follow the discussion in the rest of the book.

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-1-108-48002-4 — Fundamentals of Modern VLSI Devices
Yuan Taur , Tak H. Ning
Excerpt
More Information

Starting with the energy bands in silicon, Chapter 2 introduces the basic concepts of Fermi level, carrier concentration, drift and diffusion current transport, and Poisson's equation. Also addressed in this chapter are generation and recombination, minority carrier lifetime, and current continuity equation.

Chapter 3: p–n Junctions and Metal–Silicon Contacts

Chapter 3 covers the basic physics and operation of p−n junctions and Schottky diodes as well as metal−silicon contacts in general. p−n junctions are basic building blocks of bipolar transistors and key components of MOSFETs. Basic knowledge of their characteristics is a prerequisite to further understand the operation of bipolar devices and for designing MOSFETs. And basic knowledge of Schottky diodes is prerequisite to understanding metal−silicon contacts in general and for designing ohmic contacts with low contact resistance. The chapter ends with a discussion of high-field effects in reverse-biased diodes.

Chapter 4: MOS Capacitors

Chapter 4 covers the fundamentals of MOS capacitors – a prerequisite to MOSFET transistors. Starting with the basic concepts of free electron level and work function, the chapter proceeds to the solution of charge and potential in silicon, followed by a full description of the *C–V* characteristics. Quantum mechanical effects, important for MOS capacitors of thin oxides, are then discussed. Added in the third edition is a new section on interface states and oxide traps. Lastly, the high field section covers tunneling currents, high-κ gate dielectrics, and gate oxide reliability.

Chapter 5: MOSFETs: Long Channel

Chapter 5 describes the basic characteristics of MOSFET devices, using n-channel MOSFET as an example for most of the discussions. It deals with the more elementary long-channel MOSFETs, with sections on the charge sheet model, regional *I–V* models, and subthreshold current characteristics. A recently developed non-GCA model gives insights to the saturation region behavior while clarifying the misleading term of "pinch-off" in most standard textbooks. In the section on channel mobility, the strain effects, both biaxial and uniaxial, on electron and hole mobilities are discussed. The last section addresses the body effect, temperature effect, and quantum effect on the long-channel threshold voltage.

Chapter 6: MOSFETs: Short Channel

This chapter deals with the more complex short-channel MOSFETs. Most circuits are built with short-channel devices because of their higher current and lower capacitance. Among the main topics are short-channel effects, scale length model, velocity saturation, and non-local transport. A ballistic MOSFET model is described on the current

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE
Cambridge University Press
978-1-108-48002-4 — Fundamentals of Modern VLSI Devices
Yuan Taur , Tak H. Ning
Excerpt
More Information

limit of a MOSFET. Next considered are the major device design issues in a CMOS technology: choice of threshold voltage based on the off-current requirement and on-current performance, power supply voltage, design of nonuniform channel doping, and discrete dopant effects on threshold voltage. The last section discusses high-field effects in a short-channel MOSFET.

Chapter 7: Silicon-on-Insulator and Double-Gate MOSFETs

Chapter 7 deals with fully-depleted SOI and double-gate MOSFETs. A general, asymmetric double-gate model is applied to long channel SOI MOSFETs. For symmetric double-gate MOSFETs – the generic form of FinFETs, an analytic potential model is described that covers all regions of operation continuously. The scale length model first introduced in Chapter 6 for bulk MOSFETs is modified for short-channel DG MOSFETs. Nanowire MOSFET models, both long and short channel, are also discussed. The last section examines the scaling limits of DG and nanowire MOSFETs based on quantum mechanical considerations.

Chapter 8: CMOS Performance Factors

This chapter begins by reviewing MOSFET scaling – the guiding principle for achieving density, speed, and power improvements in VLSI evolution. The implications of the non-scaling factors, specifically, thermal voltage and silicon bandgap, on the path of CMOS evolution are discussed. The rest of the chapter deals with the key factors that govern the switching performance and power dissipation of basic digital CMOS circuits. After a brief description of static CMOS logic gates, their layout and noise margin, Section 8.3 considers the parasitic resistances and capacitances that may adversely affect the delay of a CMOS circuit. These include source and drain series resistance, junction capacitance, overlap capacitance, gate resistance, and interconnect capacitance and resistance. In Section 8.4, a delay equation is formulated and applied to study the sensitivity of CMOS delay to a variety of device and circuit parameters such as wire loading, device width and length, gate oxide thickness, power-supply voltage, threshold voltage, parasitic components, and substrate sensitivity in stacked circuits. The last section addresses the performance factors of MOSFETs in RF circuits, in particular, the unity-current-gain frequency and unity-power-gain frequency.

Chapter 9: Bipolar Devices

The basic components of a bipolar transistor are described in Chapter 9. Both vertical bipolar transistors, including SiGe-base transistors, and symmetric lateral bipolar transistors on SOI are covered. The discussion focuses on the vertical n–p–n transistors, since they are the most commonly used. The difference between n–p–n vertical transistors and symmetric lateral n–p–n transistors are pointed out where appropriate.

The basic operation of a bipolar transistor is described in terms of two p–n diodes connected back to back. The basic theory of a p–n diode is modified and applied to

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-1-108-48002-4 — Fundamentals of Modern VLSI Devices
Yuan Taur , Tak H. Ning
Excerpt
More Information

derive the current equations for a bipolar transistor. From these current equations, other important device parameters and phenomena, such as current gain, early voltage, base widening, and diffusion capacitance, are examined. The basic equivalent-circuit models relating the device parameters to circuit parameters are developed. These equivalent-circuit models form the starting point for discussing the performance of a bipolar transistor in circuit applications.

Chapter 10: Bipolar Device Design

Chapter 10 covers the basic design of a bipolar transistor. The design of the individual device regions, namely the emitter, the base, and the collector, are discussed separately. Since the detailed characteristics of a bipolar transistor depend on its operating point, the focus of this chapter is on optimizing the device design according to its intended operating condition and environment, and on the tradeoffs that must be made in the optimization process. The physics and characteristics of vertical SiGe-base bipolar transistors are discussed in depth. The design of symmetric lateral bipolar transistors on SOI is also covered, including the development of analytical models for the device parameters, base and collector currents, and the transit times.

Chapter 11: Bipolar Performance Factors

The major factors governing the performance of bipolar transistors in circuit applications are discussed in Chapter 11. Several of the commonly used figures of merit, namely, cutoff frequency, maximum oscillation frequency, and logic gate delay, are examined, and how a bipolar transistor can be optimized for a given figure of merit is discussed. Sections are devoted to examining the important delay components of a logic gate, and how these components can be minimized. The scaling properties of vertical bipolar transistors for high-speed digital logic circuits are discussed. A discussion of the optimization of bipolar transistors for RF and analog circuit applications is given. The chapter concludes with a discussion of the design tradeoff and optimization of symmetric lateral bipolar transistors for RF and analog circuit applications. Finally, several unique opportunities offered by symmetric lateral bipolar transistors, some of them beyond the capability of CMOS, are discussed.

Chapter 12: Memory Devices

In Chapter 12, the basic operational and device design principles of commonly used memory devices are discussed. The memory devices covered include CMOS SRAM, DRAM, bipolar SRAM, and several commonly used in nonvolatile memories. Typical read, write, and erase operations of the various memory arrays are explained. The issue of noise margin in scaled CMOS SRAM cells is discussed. A brief discussion of more recent developments of NAND flash technologies, including multi-bit per cell, 3D NAND, and wear leveling is given.

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-1-108-48002-4 — Fundamentals of Modern VLSI Devices
Yuan Taur , Tak H. Ning
Excerpt
More Information

2 Basic Device Physics

2.1 Energy Bands in Silicon 9
2.2 n-Type and p-Type Silicon 15
2.3 Carrier Transport in Silicon 21
2.4 Basic Equations for Device Operation 28

This chapter reviews the basic concepts of semiconductor device physics. It covers energy bands in silicon, Fermi level, n-type and p-type silicon, electrostatic potential, drift and diffusion current transport, and basic equations governing VLSI device operation. These will serve as the basis for understanding more advanced device concepts discussed in the rest of the book.

2.1 Energy Bands in Silicon

The starting material used in the fabrication of VLSI devices is silicon in the crystalline form. The silicon wafers are cut parallel to either the <111> or <100> planes (Sze, 1981), with <100> material being the most commonly used. This is largely due to the fact that <100> wafers, during processing, produce the lowest charges at the oxide–silicon interface as well as higher mobility (Balk *et al*., 1965). In a silicon crystal each atom has four valence electrons to share with its four nearest neighboring atoms. The valence electrons are shared in a paired configuration called a *covalent bond*. ***The most important result of the application of quantum mechanics to the description of electrons in a solid is that the allowed energy levels of electrons are grouped into bands*** (Kittel, 1976). ***The bands are separated by regions of energy that the electrons in the solid cannot possess: forbidden gaps***. The highest energy band that is completely filled by electrons at 0 K is called the *valence band*. The next highest energy band, separated by a forbidden gap from the valence band, is called the *conduction band*, as shown in Figure 2.1.

2.1.1 Bandgap of Silicon

What sets a semiconductor such as silicon apart from a metal or an insulator is that, at absolute zero temperature, the valence band is completely filled with electrons, while the conduction band is completely empty, and that the separation between the conduction band and valence band, or the *bandgap*, is on the order of 1 eV. On the one hand, no electrical conduction is possible at 0 K, since there are no electrons in the conduction band, whereas the electrons in the completely filled valence band cannot be accelerated by an electric field and gain energy. On the other hand, the bandgap is small enough that at room temperature a small fraction of the electrons are excited into the conduction band, leaving behind vacancies, or *holes*, in the valence

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-1-108-48002-4 — Fundamentals of Modern VLSI Devices
Yuan Taur , Tak H. Ning
Excerpt
More Information

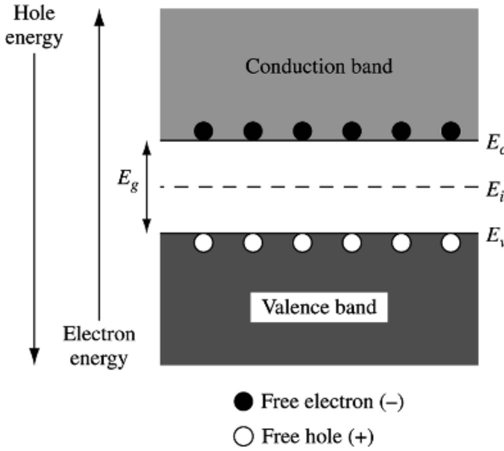

Figure 2.1 Energy-band diagram of silicon

band. This allows limited conduction to take place from the motion of both the electrons in the conduction band and the holes in the valence band. In contrast, an insulator has a much larger forbidden gap of at least several electron volts, making room-temperature conduction virtually nonexistent. Metals, on the contrary, have partially filled conduction bands even at absolute zero temperature, so that the electrons can easily move into states of higher energy in response to an applied electric field. This makes them good conductors at any temperature.

As shown in Figure 2.1, the energy of the electrons in the conduction band increases upward, while the energy of the holes in the valence band increases downward. The bottom of the conduction band is designated E_c, and the top of the valence band E_v. Their separation, or the bandgap, is $E_g = E_c - E_v$. For silicon, E_g is 1.12 eV at room temperature or 300 K. The bandgap decreases slightly as the temperature increases, with a temperature coefficient of $dE_g/dT \approx -2.73 \times 10^{-4}\text{eV/K}$ for silicon near 300 K. Other important physical parameters of silicon and silicon dioxide are listed in Table 2.1 (Green, 1990).

2.1.2 Density of States

The density of available electronic states within a certain energy range in the conduction band is determined by the number of different momentum values that can be acquired by electrons in this energy range. Based on quantum mechanics, there is one allowed state in a phase space of volume $(\Delta x\,\Delta p_x)(\Delta y\,\Delta p_y)(\Delta z\,\Delta p_z) = h^3$, where p_x, p_y, p_z are the x-, y-, z-components of the electron momentum, respectively, and h is Planck's constant. If we let $N(E)dE$ be the number of electronic states per unit volume with an energy between E and $E + dE$ in the conduction band, then

$$N(E)\,dE = 2g\frac{dp_x\,dp_y\,dp_z}{h^3}, \tag{2.1}$$

© in this web service Cambridge University Press www.cambridge.org