1 Introduction

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 thirty 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 $250 billion in 2007. 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 concept to VLSI applications in the mid 1980s, can be found in the paper by Sah (Sah, 1988). Figure 1.1 gives a chronology of the major milestone events in the development of VLSI technology. The bipolar transistor technology was developed early on and was applied to the first integrated-circuit memory in mainframe computers in the 1960s. 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 of bipolar circuits has severely limited their integration level, to about 10^4 circuits per chip. This integration level is quite low by today's VLSI standard.

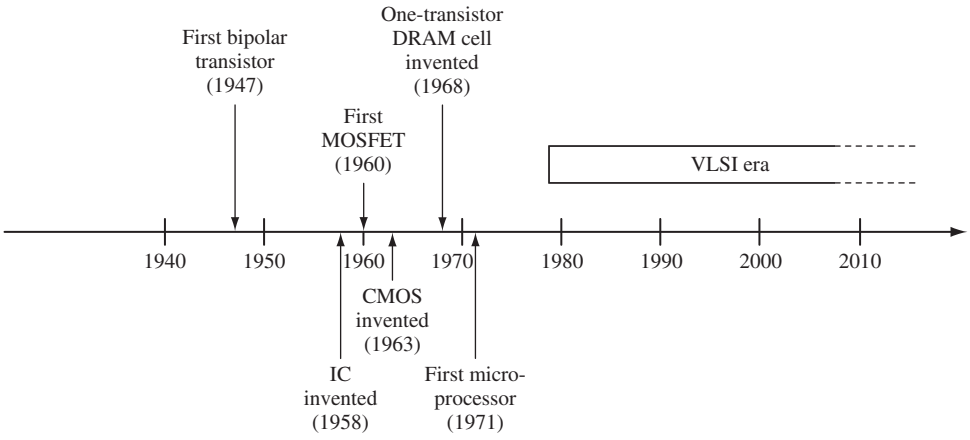

Figure 1.1. A brief chronology of the major milestones in the development of VLSI.

The idea of modulating the surface conductance of a semiconductor by the application of an electric field was first reported 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 bipolar circuits, single-polarity MOSFET circuits suffered from large standby power dissipation, and 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 hundreds of millions 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 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.

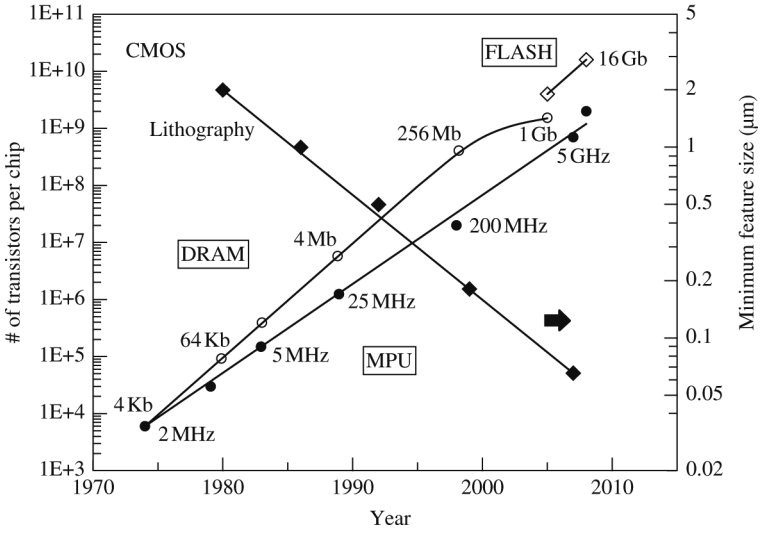

Figure 1.2. Trends in lithographic feature size, number of transistors per chip for DRAM and microprocessors (MPU), and number of memory bits per chip for Flash. The transistor count for DRAM is computed as 1.5 times the number of bits on the chip to account for the peripheral circuits. Recent data points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

As linear dimensions reached the 0.5-μm level in the early 1990s, the performance advantage of bipolar transistors was outweighed by the significantly greater circuit density of CMOS devices. The system performance benefit of integrated functionality superseded that of raw transistor performance. Even the 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 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 history and recent trends in these areas are illustrated in Fig. 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 $1.2 \times 1.2\ \text{cm}^2$.

One remarkable feature of silicon devices that fuels 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 today are 10 times faster and occupy less than 1% of the area of those built 20 years ago. This is illustrated in the trend of microprocessor units (MPUs) in Fig. 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 first edition of this book in 1998, there have been major developments in the VLSI industry that are worth mentioning. These include the following.

- Up until the mid 1990s, DRAM has been the technology driver (ITRS, 1999). However, since the mid 1990s, microprocessor has replaced DRAM as the driver of VLSI technology. This shift occurred because microprocessors push the CMOS devices to shorter gate lengths and lower supply voltages and require many more wiring levels than DRAM (ITRS, 2007). The demand in microprocessor performance has spun recent research activities in high-κ gate dielectrics as a replacement for SiO_2 and in materials and device structures with enhanced transport properties. Some of the advanced features have already shown up in selected 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. Today, NAND Flash has overtaken DRAM as the IC chip with the highest component count, as shown in Fig. 1.2 (Kim, 2008).
- Two silicon derivative technologies, SOI (silicon on insulator) CMOS and SiGe bipolar, have gone into volume manufacturing. SOI CMOS is used primarily in high-end computers and interactive game systems for additional device performance. SiGe-base bipolar, with its greatly improved frequency response and analog-circuit attributes, is used in many RF and analog circuits today.
- With the bulk CMOS devices scaled to nearing their limits, researchers in the VLSI area have been exploring double-gate MOSFETs, and in general, multiple-gate MOSFETs which in principle can extend CMOS scaling to 10 nm gate lengths and below.

1.2 Modern VLSI Devices

It is clear from Fig. 1.2 that modern transistors of practical interest have feature sizes of 0.5 μm and smaller. Although the basic operation principles of large and small transistors are the same, the relative importance of the various device parameters and performance factors for the small-dimension modern transistors is quite different from that for the transistors of the early 1980s or earlier. It is our intention to focus our discussion in this book on the fundamentals of silicon devices of sub-0.5-μm generations.

1.2.1 Modern CMOS Transistors

A schematic cross section of modern CMOS transistors, consisting of an n-channel MOSFET and a p-channel MOSFET integrated on the same chip, is shown in Fig. 1.3.

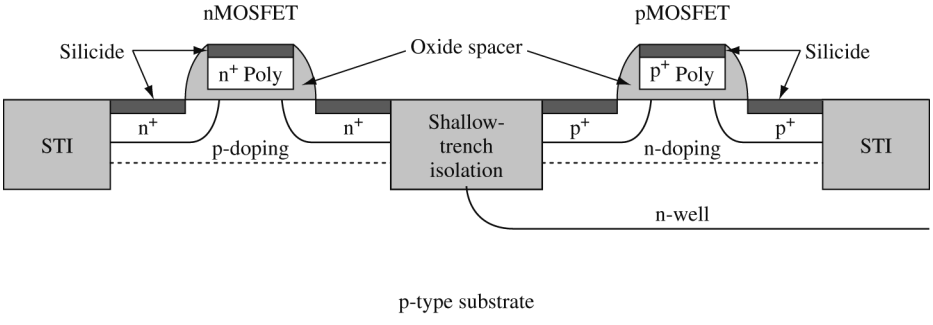

Figure 1.3. Schematic device cross section for an advanced CMOS technology.

A generic process flow for fabricating the CMOS transistors is outlined in Appendix 1. The key physical features of the modern CMOS technology, as illustrated in Fig. 1.3, include: p-type polysilicon gate for the p-channel MOSFET and n-type polysilicon gate for the n-channel MOSFET, refractory metal silicide on the polysilicon gate as well as on the source and drain diffusion regions, and shallow-trench oxide isolation.

In the electrical design of the modern CMOS transistor, the power-supply voltage is reduced with the physical dimensions in some coordinated manner. A great deal of design detail goes into decreasing the channel length, or separation between the source and drain, maximizing the on current of the transistor while maintaining an adequately low off current, minimizing variation of the transistor characteristics with process tolerances, and minimizing the parasitic resistances and parasitic capacitances.

1.2.2 Modern Bipolar Transistors

Figure 1.4 shows the schematic cross sections of two modern bipolar transistors: (a) with a Si-base and (b) with a SiGe-base. The process outline for fabricating transistor (a) is shown in Appendix 2. The salient features of the modern bipolar transistors include: shallow-trench field oxide and deep-trench isolation, polysilicon emitter, polysilicon base contact which is self-aligned to the emitter contact, and a pedestal collector which is doped to the desired level only directly underneath the emitter. A SiGe-base transistor is superior to a Si-base transistor for RF and analog circuit applications.

Unlike CMOS, the power-supply voltage for a bipolar transistor is usually kept constant as the transistor physical dimensions are reduced. Without the ability to reduce the operating voltage, electrical breakdown is a severe concern in the design of modern bipolar transistors. In designing a modern bipolar transistor, a lot of effort is spent tailoring the doping profile of the various device regions in order to maintain adequate breakdown-voltage margins while maximizing the device performance. At the same time, unlike the bipolar transistors before the early 1980s when the device performance was mostly limited by the device physical dimensions practical at the time, a modern bipolar transistor often has its performance limited by its current-density capability and

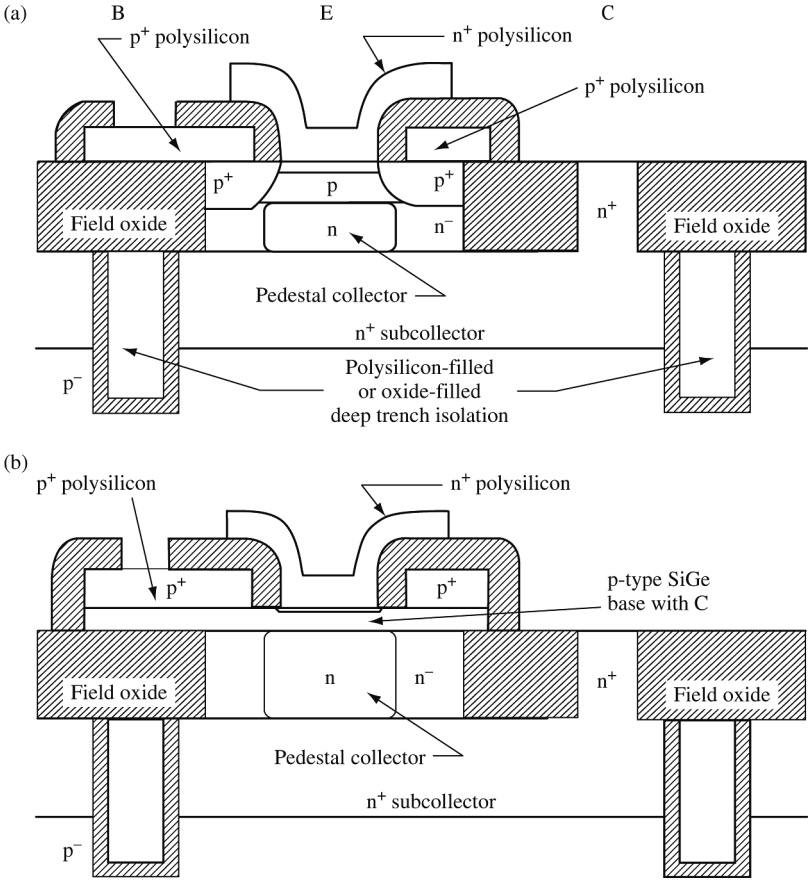

Figure 1.4. Schematic cross sections of modern silicon n–p–n bipolar transistors. (a) A transistor having a Si-base doped by ion implantation. (b) A transistor having a SiGe-base doped *in situ* with boron. Carbon is often added to suppress boron diffusion in the base layer.

not by its physical dimensions. Attempts to improve the current-density capability of a transistor usually lead to reduced breakdown voltages.

1.3 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 will be 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. Below is a brief description of each chapter. Two new chapters are added in the second edition: one on memory devices and the other on SOI devices.

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.

Starting with the energy bands in silicon, Chapter 2 first introduces the basic concepts of Fermi level, carrier concentration, drift and diffusion current transport, and Poisson's equation. The next two sections focus on the most elementary building blocks of silicon devices: the p–n junction and the MOS capacitor. Basic knowledge of their characteristics is a prerequisite to further understand the operation of the VLSI devices they lead into: bipolar and MOSFET transistors. The rest of Chapter 2 covers high-field effects, Si–SiO_2 systems, metal–silicon contacts, hot carriers, and the physics of tunneling and breakdown relevant to VLSI device reliability.

Chapter 3: MOSFET Devices

Chapter 3 describes the basic characteristics of MOSFET devices, using the n-channel MOSFET as an example for most of the discussions. It is divided into two parts. The first part deals with the more elementary long-channel MOSFETs, including subsections on drain current models, I–V characteristics, subthreshold currents, channel mobility, and intrinsic capacitances. These serve as a foundation for understanding the more important but more complex short-channel MOSFETs, which have lower capacitances and carry higher currents per gate voltage swing. The second part of Chapter 3 covers the specific features of short-channel MOSFETs important for device design purposes. The subsections include short-channel effects, velocity saturation and high-field transport, channel-length modulation, and source–drain series resistance.

Chapter 4: CMOS Device Design

Chapter 4 considers the major device design issues in a CMOS technology. It begins with the concept of MOSFET scaling – the most important guiding principle for achieving density, speed, and power improvements in VLSI evolution. Several non-scaling factors are addressed, notably, the thermal voltage and the silicon bandgap,

which have significant implications on the deviation of the CMOS evolution path from ideal scaling. Two key CMOS device design parameters – threshold voltage and channel length – are then discussed in detail. Subsections on threshold voltage include off-current requirement, choice of gate work function, channel profile design, nonuniform doping, and quantum-mechanical and discrete dopant effects on threshold voltage. Subsections on channel length include the definition of effective channel length, its extraction by the conventional method and the shift-and-ratio method, and the physical interpretation of effective channel length.

Chapter 5: CMOS Performance Factors

Chapter 5 examines the key factors that govern the switching performance and power dissipation of basic digital CMOS circuits which form the building blocks of a VLSI chip. Starting with a brief description of static CMOS logic gates, their layout and noise margin, we examine 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. Next, we formulate a delay equation and use it to study the sensitivity of CMOS delay performance 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 of Chapter 5 further extends the discussion of performance factors to several advanced CMOS materials and device structures. These include RF CMOS, effect of mobility on CMOS delay, and low-temperature CMOS.

Chapter 6: Bipolar Devices

The basic components of a bipolar transistor are described in Chapter 6. The discussion is based entirely on the vertical n–p–n transistor, since practically all high-speed bipolar transistors used in digital circuits are of the vertical n–p–n type. However, the basic device operation concept and device physics can be readily extended to other types of bipolar transistors, such as p–n–p bipolar transistors and lateral bipolar transistors.

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 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–collector junction avalanche, emitter–collector punch-through, base widening, and diffusion capacitance, are examined. Finally, 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 7: Bipolar Device Design

Chapter 7 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 sections include an examination of the effect of grading the base doping profile to enhance the drift field in the intrinsic base, and a derivation of the collector-current equations when there is significant heavy doping effect in the base. In addition, the physics and characteristics of SiGe-base bipolar transistors are discussed in much greater depth than in the first edition. The chapter concludes with a discussion of the salient features of the most commonly used modern bipolar device structure.

Chapter 8: Bipolar Performance Factors

The major factors governing the performance of bipolar transistors in circuit applications are discussed in Chapter 8. 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 power–delay tradeoffs in the design of a bipolar transistor under various circuit-loading conditions are also examined. The scaling properties of bipolar transistors, and how the large standby power dissipation of bipolar circuits limits the integration level of bipolar chips, are discussed. A discussion of the optimization of bipolar transistors for RF and analog circuit applications is given. The chapter concludes with a discussion comparing SiGe-base bipolar transistors with GaAs heterojunction bipolar transistors.

Chapter 9: Memory Devices

In Chapter 9, 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 nonvolatile memories including Flash. 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.

Chapter 10: Silicon-on-Insulator Devices

The last chapter of this book deals with silicon-on-insulator (SOI) devices, which include SOI CMOS, SOI bipolar, and double-gate MOSFETs. Both partially depleted and fully depleted SOI MOSFETs and their scaling characteristics are covered. A recently developed analytic-potential model for the drain current of a symmetric double-gate MOSFET is discussed at the end.

Appendices

There are altogether 18 appendices in the back of this book, covering in more detail various topics ranging from generation and recombination, analytic short-channel threshold model, quantum mechanical solution in weak inversion, emitter and base series resistance, to unity-gain frequencies of MOSFET and bipolar transistors. They usually involve mathematical treatments too tedious and lengthy to be included in the main text. Ten of the 18 appendices are new additions to the second edition.