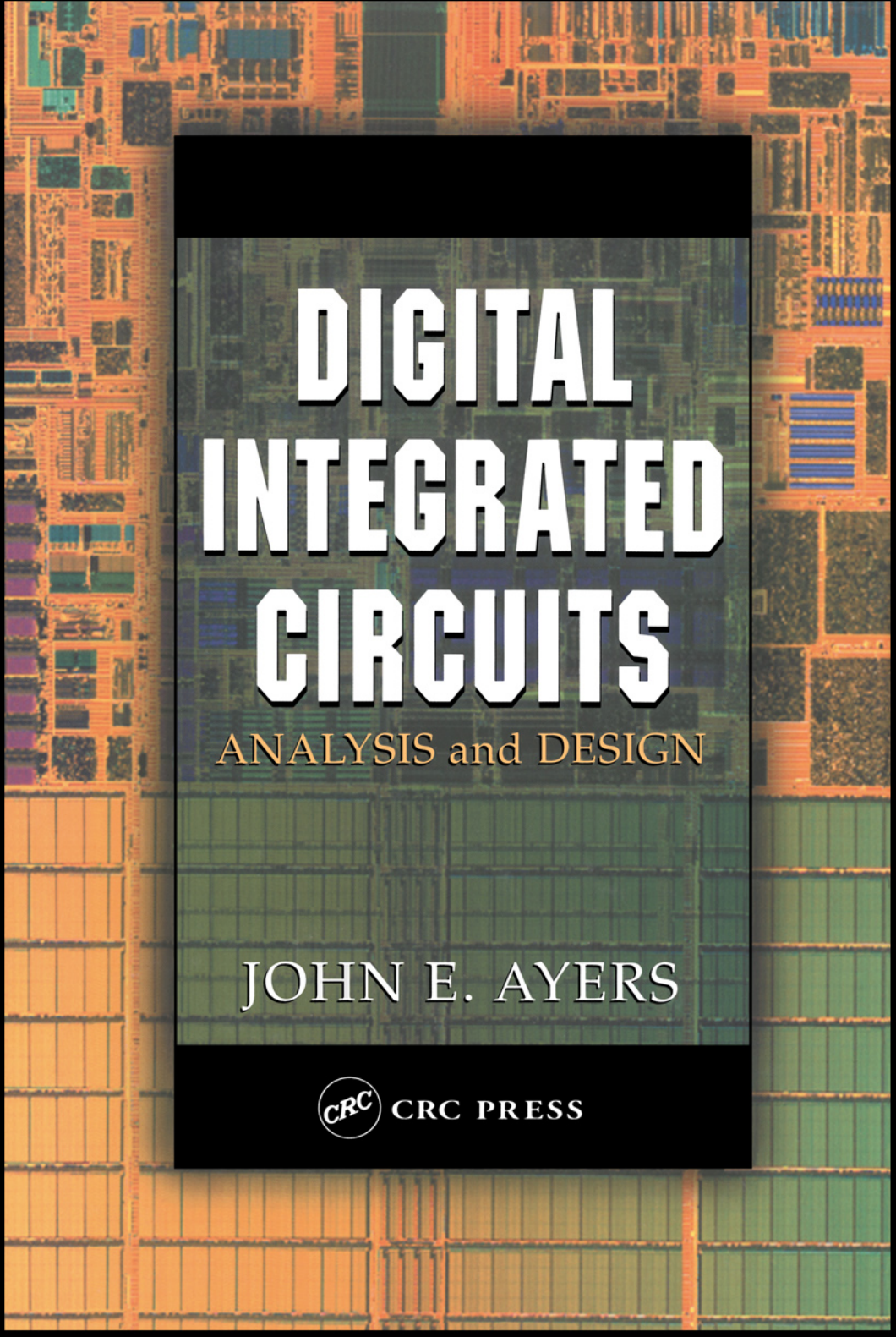


DIGITAL INTEGRATED CIRCUITS

ANALYSIS and DESIGN

JOHN E. AYERS
University of Connecticut

CRC PRESS

Boca Raton London New York Washington, D.C.

Cover photo: Die shot of the Intel® Pentium® M Processor, part of Intel® Centrino™ Mobile Technology. (Photo courtesy of Intel.)

This edition published in the Taylor & Francis e-Library, 2005.

Library of Congress Cataloging-in-Publication Data

Ayers, J. E. (John E.)
Digital integrated circuits : analysis and design / J.E. Ayers.
p. cm.
Includes bibliographical references and index.
ISBN 0-8493-1951-X (alk. paper)
1. Digital integrated circuits—Design and construction. I. Title.

TK7874.65.A94 2003
621.3815—dc22 2003055586

International Standard Book Number 0-8493-1951-X
Library of Congress Card Number 2003055586

ISBN 0-203-48690-0 Master e-book ISBN

ISBN 0-203-58907-6 (Adobe eReader Format)

To Ruth Bridges Ayers, researcher and author;

to George H. Ayers, Jr., scientist and teacher;

and to Kimberly, Jacob, Sarah, and Rachel, for making it all worthwhile.

Preface

No field of enterprise today is more dynamic or challenging than that of digital integrated circuits. Since the invention of the integrated circuit in 1958, our ability to pack transistors on a single chip of silicon has doubled roughly every 18 months, as described by "Moore's law." As a consequence, the functionality and performance of digital integrated circuits have improved geometrically with time. This exponential progress is unprecedented in any other industry or segment of the world economy, and has revolutionized the way we live and work.

Because of its very nature, the field of digital integrated circuits has rapidly outrun the numerous good books available on the topic. In response, some authors have adopted the approach of narrowing the focus to a single subfield, with the goal of covering an ever-increasing wealth of technology. None, however, has made a clear transition to the modern multidisciplinary practice of digital integrated circuits.

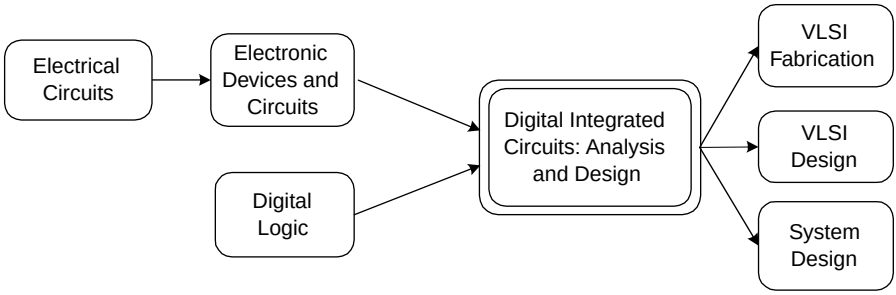
Traditionally, engineers at the materials, process, device, circuit, and system levels worked quite separately. VLSI design rules developed by Mead and Conway freed the circuit designer from the need to understand the details of device design or fabrication. Rapid progress in scaling transistor dimensions has rendered it impossible to compartmentalize our expertise in this way, however. Engineers working in the field of digital integrated circuits must understand materials, physics, devices, processing, electromagnetics, computer tools, and economics, as well as circuits and layout design rules.

Recent innovations in interconnect, such as copper and low k dielectrics, came about by the application of materials, processing, circuit, and electromagnetics principles. The emergence of silicon-on-insulator (SOI) resulted from the application of materials, processing, and device physics as well as circuit theory. At the same time, yield and economic issues have guided the course of SOI development to where it is today. Successful implementation of a system on chip (SOC) can be done only with an understanding of process, yield, economic, and packaging trade-offs. Emerging memory technologies have benefited from interdisciplinary work in physics, materials, and devices. The interdisciplinary nature of the field is highlighted by ovonic unified memory (OUM), which borrows materials technology from rewritable compact disks.

Digital Integrated Circuits: Analysis and Design was created with three goals:

1. To present an interdisciplinary approach that will remain relevant for years to come
2. To provide broad coverage of the field that is relevant for engineers designing integrated circuits or *designing with* integrated circuits
3. To focus on the underlying principles, rather than on details of current technology that will soon be obsolete

The approach of this book will render it useful for students and practicing engineers alike. Undergraduates and graduates in many fields, including computer engineering, electrical engineering, computer science, materials, physics, and manufacturing will benefit from this book. In a modern electrical and computer-engineering curriculum, *Digital Integrated Circuits: Analysis and Design* fits into the junior or senior year as shown schematically here.



Students taking the course are assumed to have a core engineering and science background, including calculus, differential equations, physics, and chemistry, as well as courses in circuits, electronics, and digital logic. The content of this book will prepare engineers for a follow-up course in very large-scale integrated circuit (VLSI) design, with an understanding of

- Digital circuits and their performance attributes and trade-offs
- Device and interconnect characteristics and design
- Circuit fabrication and associated design rules
- Computer simulation

Similarly, an understanding of the interplay among materials, processing, device, and circuit issues will serve as the groundwork for a VLSI fabrication course. Although some electrical and computer engineers will design digital integrated circuits, all electrical and computer engineers will be involved in *design with* digital integrated circuits. These engineers will be prepared with an understanding of the principles of digital circuits and their attributes, including bipolar, MESFET, MOS, and BiCMOS circuits, their manufacture, testing, and reliability, interfacing, and packaging.

Chapter Overview

Chapter 1 provides an overview of the interdisciplinary field of digital integrated circuits, including issues of economics (Moore's law, the international roadmap for semiconductor technology, circuit yield), circuits (logic function, electrical performance attributes and trade-offs), computer tools, fabrication (bipolar, MESFET, and MOSFET circuits, and silicon-on-insulator), reliability, burn-in and testing.

Chapter 2 reviews the basic semiconductor materials and physics necessary for an understanding of devices.

Chapter 3 and Chapter 4 describe bipolar devices, their basic physics, fabrication, and computer models.

Chapter 5 and Chapter 6 cover saturated and current-mode bipolar logic circuits, including transistor–transistor logic (TTL) and important variations, and emitter-coupled logic (ECL). In each case, the circuit evolution is described to promote an understanding of the subtle design features in more complex circuit versions. High-performance circuit techniques, such as active pull-down ECL, low-voltage circuits, and advanced Schottky design concepts for TTL, are presented.

Chapter 7 provides a firm grounding in the physics and models for field-effect transistors, with an emphasis on the metal oxide–semiconductor field-effect transistor (MOSFET). The principles of subthreshold operation, the body bias effect, and short-channel MOSFET operations are discussed.

Chapter 8 through Chapter 10 cover the principles of MOS logic circuits, including NMOS and CMOS, dynamic logic gates, and their modeling. The importance of low-power CMOS design warranted the creation of its own chapter. Chapter 10 presents important low-power CMOS design concepts and trade-offs, including low-voltage CMOS, multiple threshold CMOS, and adiabatic logic. The interdisciplinary nature of low-power CMOS design is evident in active body biasing and silicon-on-insulator (SOI) for low-power CMOS.

Chapter 11 presents the principles of bipolar–CMOS (BiCMOS) logic circuits and the trade-offs in logic swing, speed, and power governing the use of BiCMOS vs. CMOS circuits.

Chapter 12 provides a firm grounding in MESFET-based logic circuits, with the focus on gallium arsenide direct-coupled FET logic (DCFL). This chapter spans the topic, including MESFET physics and models, DCFL circuits and models, and computer simulation.

Chapter 13 addresses the principles of interfacing, including level-shifting circuits, wired logic, transmission gates, and tri-state logic.

Chapter 14 provides a firm grounding in interconnect, including interconnect parasitics, circuit and transmission line models, materials issues such as copper and low-k dielectrics, and special problems in interconnect design.

Chapter 15 addresses the concepts underlying bistable circuits, including latches, flip-flops, and Schmitt triggers.

Chapter 16 presents the principles of digital memories, including their organization, basic circuit types and attributes, and their application. Access times are discussed from the viewpoint of interconnect delays and emerging memory concepts are presented.

Chapter 17 introduces the principles of physical integrated circuit design and makes connections among fabrication technology, lithography, and design rules. Design rules for MOSFETs, bipolar devices, and passive components are presented. Example device and circuit designs are provided and VLSI design is introduced.

Chapter 18 presents the principles underlying integrated circuit packages, including electrical, chemical, mechanical, and thermal issues.

The underlying philosophy of the book has been to focus on principles, to include bipolar as well as MOS concepts, and to present everything from a modern interdisciplinary point of view. Computer tools are stressed throughout, and nearly every chapter includes SPICE* examples and exercises. Most chapters include laboratory exercises, to enable use with modern courses integrating lecture and laboratory components into a single offering. Every chapter is followed by a "quick reference" that collects the important concepts, diagrams, equations, and design rules together for convenient access. It is hoped that these attributes will make *Digital Integrated Circuits: Analysis and Design* valuable not only to engineers designing integrated circuits, but also to the larger number of engineers designing with digital integrated circuits.

I would like to thank my colleagues for their encouragement and illuminating discussions and give my heartfelt thanks to S.K. Ghandhi for his guidance and inspiration. Over the past 3 years my students have been subjected to numerous versions of this manuscript; to them I owe my gratitude. Special thanks are due to Jeff Allanach at the University of Connecticut for his assistance in preparing this manuscript.

J.E. Ayers
Storrs, Connecticut

---

* SPICE stands for Simulation Program with Integrated Circuit Emphasis. There are many versions of SPICE in use by industry; however, fluency with one version can be adapted easily to any other. PSPICE was used for all examples in this book because a student version can be downloaded for free from www.cadence.com.

About the Author

J.E. Ayers grew up 8 miles from an integrated circuit design and fabrication facility, where he worked as a technician and first developed his passion for the topic. After earning a B.S. EE degree with highest distinction from the University of Maine, Orono, in 1984, he worked as an integrated circuit test engineer for National Semiconductor, South Portland, Maine. Following that, Dr. Ayers worked for 6 years on semiconductor material growth and characterization at Rensselaer Polytechnic Institute, Troy, New York, and Philips Laboratories, Briarcliff, New York. He earned the M.S. EE in 1987 and the Ph.D. EE in 1990, both from Rensselaer Polytechnic Institute. Since then he has been employed in academic research and teaching at the University of Connecticut, Storrs.

Dr. Ayers has authored over 40 scientific journal papers and refereed conference proceedings. He is a member of the Institute of Electrical and Electronics Engineers, the Materials Research Society, Eta Kappa Nu, Tau Beta Pi, and Phi Kappa Phi. He currently lives in Ashford, Connecticut, with his wife and three children.

Contents

1

Introduction to Digital Integrated Circuits

1.1 The Technological Revolution

The 20th century brought about an explosion of electronics technology that drastically changed the way we live and work today. The silicon era began with the invention of the bipolar transistor in 1947[1–4] at Bell Laboratories (Figure 1.1). The string of important developments that followed has led to today's gigahertz microprocessors and gigabit memories. Replacing large and bulky vacuum tubes, the transistor made it possible to build practical computers. Equally important was the invention of the integrated circuit in 1958[5,6] and subsequent improvements on the concept in the 1960s. These breakthroughs allowed the fabrication of many devices in a single chip of silicon, enabling computing power far beyond that achievable by wiring together discrete transistors.

Another significant development was the first metal oxide–semiconductor field-effect transistor (MOSFET). Even though this device was invented in 1930 by Lilienfeld, the first working MOSFET was demonstrated in 1960 by Kahng and Atalla.[7–9] Although bipolar transistors are superior to MOSFETs in raw speed, the relatively high power consumption has limited their level of integration to about 10,000 gates per chip. The small size and low power requirements of MOSFETs have greatly aided the development of complex microprocessors, high-density memory chips, mobile computers, digital cellular telephones, and many other electronic products. The first microprocessor was implemented in 1971 using MOSFETs. Complementary MOSFET (CMOS) logic, invented in 1963, is the basis for nearly all modern microprocessors. The one-transistor dynamic random access memory (DRAM) cell, invented in 1968, uses a single MOSFET for each bit and is the basis for the gigabit DRAM chip.

The key transistor and integrated circuit inventions were followed by less heralded but equally important developments that have brought about steady progress in digital integrated circuits. Soon after the realization of integrated circuits, Intel co-founder Gordon Moore noted that the number of transistors per chip was increasing exponentially with time. "Moore's law" states that the number of transistors per chip doubles every

FIGURE 1.1
The first transistor, invented at Bell Laboratories in 1947. (Courtesy of Lucent Technologies Inc.)

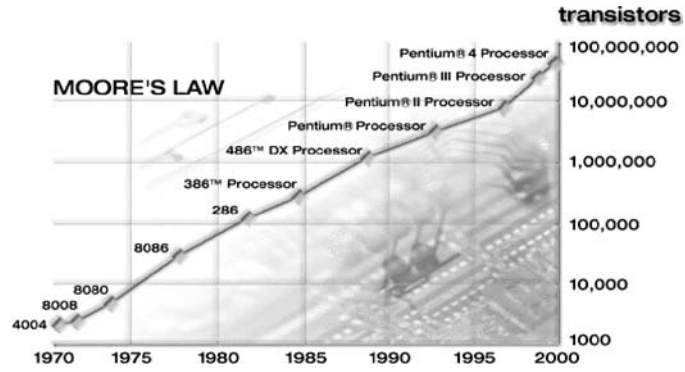


FIGURE 1.2
Trend in the number of transistors per chip for microprocessors. (Courtesy of Intel Corporation.)

18 months.[10,11] Remarkably, this rate of progress has been maintained for over three decades. Figure 1.2 illustrates this exponential progress in the case of microprocessors; similar trends have been established in dynamic random access memories (DRAMs) and application-specific integrated circuits (ASICs).

Industry has kept pace with Moore's law by two means: using ever increasing die sizes and scaling down the dimensions of transistors through improved lithography.[12,13] The first might seem trivial but is not. Increasing die sizes has required nearly flawless manufacturing processes in order to

maintain acceptable yields because an increase in the chip area is accompanied by an increased probability of a defect. The scaling of transistors has been pursued relentlessly and has brought about improvements in circuit performance and cost as well as density.

With the goal of extending the historic trends in integrated circuit technology, the Semiconductor Industry Association (SIA) in the U.S.[14] produced the National Technology Roadmap for Semiconductors (NTRS) in 1992. This roadmap defined industry-wide technology goals for a 15-year period and was revised in 1994 and 1997. In 1998, following the globalization of the semiconductor industry, an international technology roadmap for semiconductors (ITRS) was developed with participation from the semiconductor industries in Europe, Japan, Korea, and Taiwan.[15–17]

What will the digital integrated circuit industry look like in 2016? According to the 2001 ITRS, silicon wafers will grow to 450 mm in diameter while transistor gate lengths will diminish to 9 nm. As a consequence of these developments, it will be possible to buy a 28.8-GHz processor with 3 billion transistors and 4700 pins for less than one microcent per transistor! These and other important trends are charted in Table 1.1.

The exponential progress in capability and speed is unique to the electronics industry and has made it the world's most dynamic field of enterprise. Ever improving circuit densities and switching speeds, coupled with decreasing costs, have enabled development of new products and therefore new markets for electronics. These include consumer products such as digital cameras and camcorders, high-definition televisions, digital versatile disk players, digital wireless phones, digital voicemail machines, video games, and palmtop computing devices, to name a few. Many other less visible applications exist in virtually every other sector of the economy, including the telecommunications, automobile, power, food, health care, clothing, aerospace, and defense industries. The market for integrated circuits is expected to surpass 120 billion units by 2005, as shown in Figure 1.3.

The rapid developments in digital integrated circuits have revolutionized the way we live. At the same time, they have also brought about a revolution in the practice of microelectronics. The increasing complexity and shrinking dimensions in digital integrated circuits have mandated the use of sophisticated computer tools for design and analysis. These tools augment but do not replace the skills of the design engineer. Rather, good design requires the combination of computer tools with a firm grounding in the underlying principles.

1.2 Electrical Properties of Digital Integrated Circuits

In digital circuitry, signals take on one of two (or possibly more) discrete levels. This contrasts with the case of analog circuits and systems, in which

TABLE 1.1

Semiconductor Technology Trends

Year of Production	2001	2004	2007	2010	2013	2016
Lithography						
DRAM 1/2 pitch (nm)[a]	130	90	65	45	32	22
MPU/ASIC 1/2 pitch (nm)[a]	150	90	65	45	32	22
MPU printed gate length (nm)	90	53	35	25	18	13
MPU physical gate length (nm)	65	37	25	18	13	9
Microprocessor unit (MPU) characteristics						
MPU transistors per chip (millions)	97	193	386	773	1546	3092
MPU chip size (mm^2)	140	140	140	280	280	280
MPU cost (microcents per transistor)	176	62	22	7.8	2.75	0.97
MPU total package pins	1200	1600	2140	2782	3616	4702
Clock frequency (GHz)	1.684	3.99	6.74	11.51	19.35	28.8
Dynamic random access memory (DRAM) characteristics						
DRAM bits per chip (billions)	0.54	1.07	4.29	8.59	34.4	68.7
DRAM chip size (mm^2)	127	93	183	181	239	238
DRAM cost (microcents per bit)	7.7	2.7	0.96	0.34	0.12	0.042
Application-specific integrated circuit (ASIC) characteristics						
ASIC package pins	1700	2263	3012	4009	5335	7100
General						
On-chip clock frequency (GHz)	1.684	3.99	6.74	11.51	19.35	28.8
Off-chip frequency (GHz)[b]	1.684	3.99	6.74	11.51	19.35	28.8
Supply voltage (V)	1.1	1.0	0.7	0.6	0.5	0.4
Chip power dissipation (W)	130	160	190	218	251	288
Silicon wafer diameter (mm)	300	300	300	300	450	450

[a] The half pitch is defined as one half of the center-to-center distance for two wires defined on the chip surface.

[b] It is expected that a small fraction of pins will achieve a frequency equal to the internal clock frequency while most pins will achieve much lower frequencies.

Source: The 2001 International Technology Roadmap for Semiconductors.

signals can take on any value in a continuous range. In the binary digital systems commonly in use today, signals exist as sequences of ones and zeroes. The advantage of digitizing analog signals is that they can be stored, duplicated, and transmitted repeatedly without any loss in quality.

Digital circuits employ semiconductor electronic devices to process or combine binary signals in a desired fashion. These digital circuits are called logic gates and, in practice, the two binary values are represented by two distinct voltage levels. Digital integrated circuits involve the fabrication of many different electronic devices in one chip of silicon. The level of integration is classified according to the number of gates integrated on a single chip. The

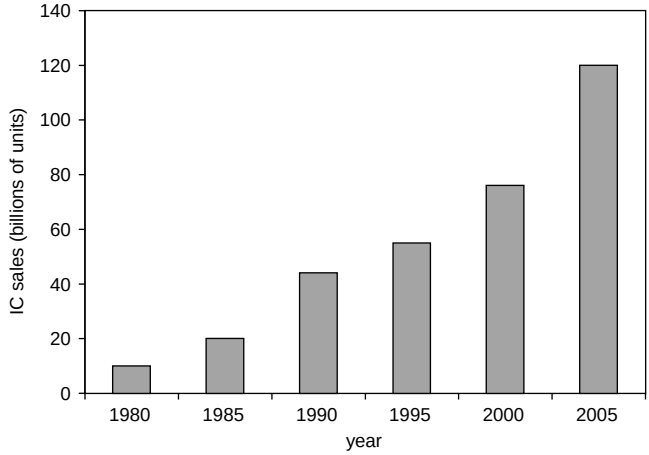


FIGURE 1.3
Trend in the worldwide integrated circuit market. (Courtesy of Semiconductor Industry Assoc.)

TABLE 1.2

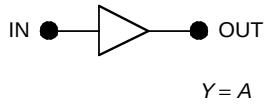
Levels of Integration

Level of Integration		**Gates/chip**
Small-scale integration	SSI	1–10
Medium-scale integration	MSI	10–100
Large-scale integration	LSI	100–10^4
Very large-scale integration	VLSI	>10^4

various levels of integration have been called small-scale integration (SSI), medium-scale integration (MSI), large-scale integration (LSI), and very large-scale integration (VLSI) and are listed in Table 1.2. This is currently the VLSI era, although all four levels of integration are in use for various applications.

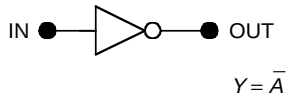
Another level of integration, called "wafer-scale integration," was proposed some years ago. The idea was to fabricate a single integrated circuit using an entire silicon wafer. This goal turned out to be far too ambitious as the size of silicon wafers grew to 200 and then 300 mm. Nonetheless, it has become feasible to implement "system on a chip" designs in which an entire computer system is built in a single chip of silicon. This approach is superior in size, cost, and performance to the traditional approach of wiring together many integrated circuits on a printed circuit board.

This section describes the electrical properties of digital integrated circuits at the gate level. Ideally, a logic gate should process an infinite number of inputs, perform some logic function with zero time delay, be completely immune to the effects of loading by other gates, and consume zero power. Although this goal has not been achieved, it serves as a starting point for the discussion of real logic gates.



IN	OUT
0	0
1	1

FIGURE 1.4
Buffer.



IN	OUT
0	1
1	0

FIGURE 1.5
Inverter.

1.2.1 Logic Function

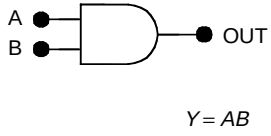
To be useful, a logic gate must perform some Boolean logic function. Boolean algebra, named after mathematician George Boole, is a system of mathematics based on the binary number system,[18] which is based on powers of two. Each binary digit, or bit, takes on a value of "0" or "1," sometimes referred to as "false" and "true" results, respectively. A string of four bits is referred to as a four-bit word, or a *nibble*. An eight-bit word is called a *byte*.

The simplest gate is a buffer, shown in Figure 1.4 along with its truth table. The value of the output Y equals the value of the input A. Although the buffer does not perform any logic function in the usual sense, it can provide conditioning of the electrical signals. For example, the buffer may provide current gain.

The other one-input logic gate is the inverter, or NOT gate, shown in Figure 1.5. If the input A is true, then the output Y is not true, and vice versa. Inversion is indicated in the Boolean equation by a bar over the inverted value. The equation shown in Figure 1.5 is read "Y equals not A." In the symbol for the inverter, inversion is shown by a circle at the output.

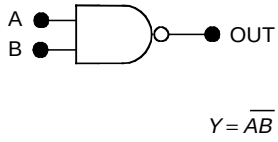
There are several important logic gates that combine two or more inputs to create the desired Boolean logic function. These include the AND, NAND, OR, NOR, and XOR gates:

- The AND gate performs the Boolean AND function of two or more inputs. For the two-input version shown in Figure 1.6, the output Y is true if and only if inputs A and B are true. This results in the truth table shown. In the Boolean algebraic equations, ANDing is shown in one of two ways: with a dot or with no symbol at all, as shown in Figure 1.6.
- The NAND function is simply an inverted version of the AND function. Figure 1.7 shows the two-input version. For this case, the output is false if and only if A and B are true; therefore the NAND



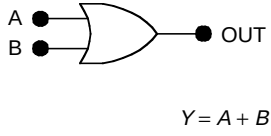
A	B	OUT
0	0	0
0	1	0
1	0	0
1	1	1

FIGURE 1.6
Two-input AND (AND2) gate.



A	B	OUT
0	0	1
0	1	1
1	0	1
1	1	0

FIGURE 1.7
NAND2 gate.

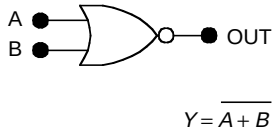


A	B	OUT
0	0	0
0	1	1
1	0	1
1	1	1

FIGURE 1.8
OR2 gate.

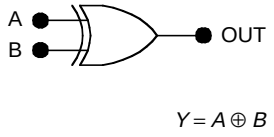
gate performs the same function as an AND gate followed by a NOT gate. As with the inverter, an overbar shows inversion of the logic function. The equation is read "Y equals NOT A AND B," or "Y equals A NAND B." Inversion is shown in the logic symbol by a small circle at the output.

- Another basic function of great importance is OR. A two-input OR gate is shown in Figure 1.8 with its truth table. For this case of two inputs, the output Y is true if either input is true. ORing is shown symbolically with a plus sign. The logic symbol is concave on the left side and pointed on the right side so that it is easily distinguished from the AND gate.
- Inversion of the OR function results in the NOR function. (NOR is short for NOT OR.) The two-input NOR gate is shown in Figure 1.9. In the Boolean equation, the NOR function is written by placing a bar over the ORed quantity. In logic diagrams, a small circle at the output symbolizes inversion.
- A logic function of great importance in adders is the exclusive OR function, abbreviated as XOR. Figure 1.10 shows the two-input version. In equations, the XOR function is represented by a plus sign



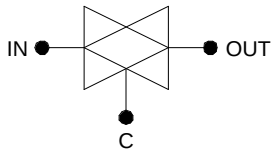
A	B	OUT
0	0	1
0	1	0
1	0	0
1	1	0

FIGURE 1.9
NOR2 gate.



A	B	OUT
0	0	0
0	1	1
1	0	1
1	1	0

FIGURE 1.10
XOR2 gate.



C	IN	OUT
0	0	High Z
0	1	High Z
1	0	0
1	1	1

FIGURE 1.11
Transmission gate.

with a circle around it. The logic symbol is similar to that for an OR gate but has a double arc on the left side. For the two-output XOR gate, the output Y is true if and only if one of the two inputs is true. The output is false if both inputs are true, distinguishing this function from OR. In terms of the NOT, OR, and AND functions, the XOR function can be written as follows:

$$Y = A \oplus B = \overline{A}B + A\overline{B}\,. \tag{1.1}$$

Another logic circuit of special importance is the transmission gate, shown in Figure 1.11. This gate is designed so that the output follows the input, as long as the control input C is at logic one. If logic zero is applied to the control input, the gate is disabled and the output is in the high-impedance (high Z) state regardless of the value of the input. With the output in the high Z state, the voltage at the output will float to whatever voltage is imposed by other circuitry connected to the node. Therefore, transmission gates can be used to connect or disconnect logic blocks in a system. This is useful in bus-based systems and power-managed digital systems.

Practical digital systems involve complex functions of many inputs that can be realized using the basic functions described previously. In fact, it is possible to realize any arbitrary logic function of any arbitrary number of inputs using only the NOT and OR functions or only the NOT and AND functions. A number of techniques are available for the simplification of complex logic functions that make it possible to realize the necessary logic functions with maximum efficiency. These techniques are beyond the scope of this book.

1.2.2 Voltage Transfer Characteristics

An important electrical characteristic of any logic gate is the *voltage transfer characteristic* (VTC). This is the output voltage vs. input voltage characteristic. It is usually measured under low-frequency, quasi-static conditions and is referred to as the DC voltage transfer characteristic. The important features of the VTC can be seen in Figure 1.12, which is a generic characteristic for an inverter. The four critical voltages for the inverter are V_{OL}, V_{OH}, V_{IL}, and V_{IH}.

The *output low voltage* (V_{OL}) is the voltage output corresponding to logic zero (a false output). The *output high voltage* (V_{OH}) is the value of the output corresponding to logic one (a true output). The difference between the two output levels is called the *logic swing*:

$$\text{LS} = V_{OH} - V_{OL}. \tag{1.2}$$

The *input low voltage* (V_{IL}) is the maximum input voltage that will be interpreted as logic zero and the *input high voltage* (V_{IH}) is the minimum

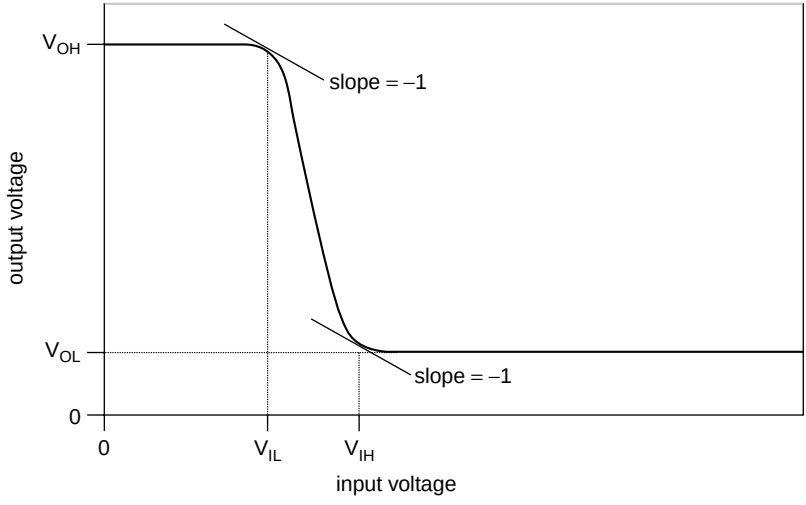


FIGURE 1.12
Voltage transfer characteristic for an inverter.

value that will be interpreted as logic one. Input values between V_{IL} and V_{IH} are ambiguous and should be avoided, so it is desirable to minimize this ambiguous range. By definition, V_{IL} and V_{IH} are the input voltages for which the slope of the transfer characteristic is –1 (+1 for a noninverting gate):

$$\left.\frac{dV_{\text{OUT}}}{dV_{\text{IN}}}\right|_{V_{IN}=V_{IL}} = -1 \tag{1.3}$$

and

$$\left.\frac{dV_{\text{OUT}}}{dV_{\text{IN}}}\right|_{V_{IN}=V_{IH}} = -1\,. \tag{1.4}$$

The *noise margins*[19] are important with regard to bit error rates in the presence of electrical noise. They are defined by

$$V_{NML} = V_{IL} - V_{OL} \tag{1.5}$$

and

$$V_{NMH} = V_{OH} - V_{IH}\,, \tag{1.6}$$

where V_{NML} and V_{NMH} are the low noise margin and the high noise margin, respectively. Electrical noise with a peak-to-peak amplitude less than the noise margin is attenuated, whereas noise of greater amplitude can create a bit error. It is therefore desirable to maximize the noise margins.

Voltage transfer characteristics are measured using the x–y feature of a storage oscilloscope or a computer-based virtual instrument. The measurement is straightforward in the case of a buffer or inverter; however, the situation is more complicated with multiple inputs. The usual approach is to tie all but one input to logic zero or logic one. This avoids the need for a multidimensional plot. For example, for a NAND gate, all inputs are tied to the positive supply voltage except one; the transfer characteristic is measured for this one input. Usually it is assumed that all inputs behave in identical fashion. For a NOR gate, all inputs but one are grounded for the measurement.

For an ideal logic gate, the output high voltage is equal to the positive supply voltage and the output low voltage is equal to the negative supply voltage (usually zero). This results in the maximum possible logic swing, called "rail to rail." The ideal logic gate also exhibits a voltage transfer characteristic with an abrupt transition midway between the supply voltages. This maximizes the two noise margins.

1.2.3 Fan-In and Fan-Out

Fan-in and fan-out refer to the connectivity of a logic gate. Fan-in is simply the number of input connections. Fan-out (or maximum fan-out) is the maximum number of load gates that can be connected to the output.

The fan-in may be unity, as in the case of an inverter; however, general system design requires gates with at least two inputs. Gates with higher fan-in are desirable because they can simplify the implementation and improve the overall performance of complex systems. Practical limits to the fan-in are imposed by device or circuit constraints. However, gates with a fan-in of eight are readily achieved in any logic family.

The maximum fan-out (N_{MAX}) is always an integer; it may be limited by static (DC) constraints or by dynamic considerations. The maximum fan-out is usually calculated with the assumption that the load gates are identical to the driving gate. The DC fan-out consideration is based on current loading. Suppose that I_{OL} is the maximum current that can be sunk at the output with a logic-zero output (the output low current). If I_{IL} is the current flowing out of an input lead when logic zero is applied (the input low current), then

$$N_{MAX} < \frac{I_{OL}}{I_{IL}} . \tag{1.7}$$

Suppose also that I_{OH} is the maximum current that can be sourced with a logic-one output (the output high current). If I_{IH} is the amount of current sunk by an input with logic one applied, then

$$N_{MAX} < \frac{I_{OH}}{I_{IH}} . \tag{1.8}$$

Typically, because the constraints imposed by Equation 1.7 and Equation 1.8 are very different, the more stringent one prevails.

The dynamic consideration is imposed by switching speed requirements. Increasing the number of load gates decreases the switching speed of the driving gate. As a consequence, there is a maximum number of load gates that can be connected without an unacceptable degradation in the switching speed. In practice, the DC and dynamic fan-out limitations are very different for a particular logic gate design; therefore, the more stringent limitation prevails. It is desirable to achieve the greatest possible value for the maximum fan-out because this makes chip design more flexible and efficient. Generally, values greater than 20 are entirely acceptable, whereas values less than 10 intrude on the chip design. These numbers are subject to the nature of the chip design, however, with more complex designs mandating higher values of the maximum fan-out.

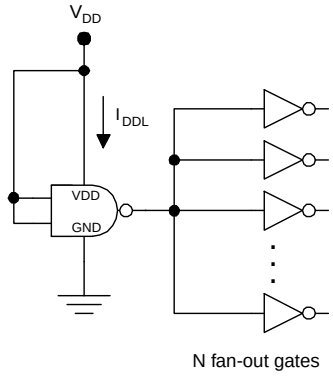


FIGURE 1.13
Determination of P_L (the output low power).

1.2.4 Dissipation

Power dissipation is an important consideration for nearly all applications of digital integrated circuits. In portable devices, the power dissipation must be minimized to prolong battery life. For all digital equipment, portable or stationary, the power dissipation must be minimized because of the associated heat that must be removed. Cooling integrated circuits often requires specially designed packages with heat sinking and fans for more efficient heat removal. In some cases, water cooling is necessary. For VLSI, dissipation considerations often limit the number of gates that can be put on a chip.

Dissipation may be dominated by the static component or by the dynamic contribution. In some cases, both contributions may be similar in magnitude, so both must be considered. The DC (static) dissipation is dependent on the output state and the fan-out.*

P_L is the DC dissipation with a low (logic-zero) output state. Referring to Figure 1.13, the output low power is

$$P_L = V_{DD} I_{DDL}\,, \tag{1.9}$$

where V_{DD} is the supply voltage and I_{DDL} is the supply current that flows with a low output. For the case shown, the gate under consideration is a two-input NAND gate. Therefore, the output low condition exists with both inputs tied to the supply voltage.

P_H is the DC dissipation for the output high state. Referring to Figure 1.14, the output high power is

$$P_H = V_{DD} I_{DDH}\,, \tag{1.10}$$

* Here "fan-out" refers to the actual number of load gates and is abbreviated N. It is always less than the maximum fan-out N_{MAX}.

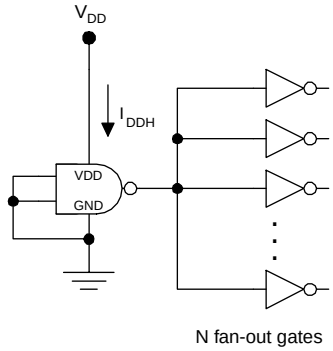


FIGURE 1.14
Determination of P_H (the output high power).

where I_{DDH} is the current flowing from the supply to the gate under the condition of a logic-one output. For the two-input NAND gate shown in the figure, the output high condition exists with both inputs tied to ground. In general, I_{DDH} and P_H are a function of N, the number of fan-out gates.

The average *DC* dissipation can be calculated by

$$P_{DC} = \frac{P_L + P_H}{2}, \tag{1.11}$$

assuming a 50% duty cycle for the output. Of course, it is not possible for all signals in a complex system to exhibit 50% duty. On the other hand, the preceding calculation represents a best estimate and provides a means for the comparison of different logic circuit designs.

The dynamic power is dissipated when the output is toggled with a capacitive load. It is given by

$$P_{AC} = fC_L V_{DD}^2, \tag{1.12}$$

where f is the switching (toggling) frequency, C_L is the load capacitance, and V_{DD} is the supply voltage. The total dissipation is the sum of the *DC* and dynamic components. Therefore,

$$P = P_{DC} + fC_L V_{DD}^2. \tag{1.13}$$

The relative contributions of the *DC* and dynamic components vary greatly among the different families of logic circuits.

1.2.5 Transient Characteristics

Transient characteristics are of great importance because of their direct relationship to the speed of digital circuits. Logic gates with improved transient

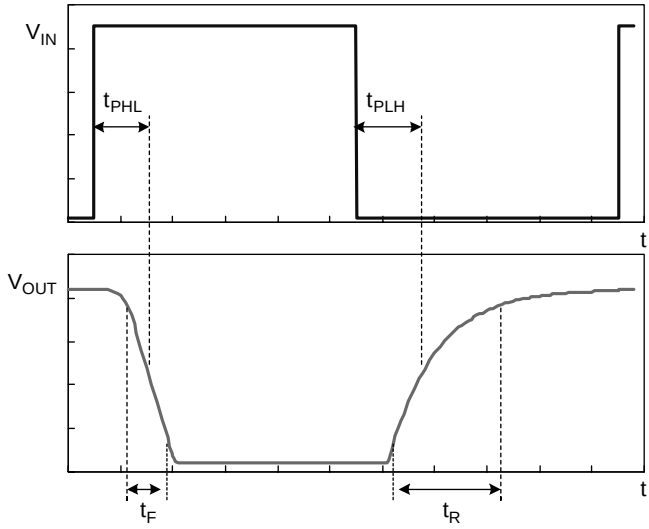


FIGURE 1.15
Transient response for an inverter with a rectangular input waveform.

response allow the use of higher clock frequencies and higher data rates. Consider Figure 1.15, which shows the transient behavior for an inverter with a rectangular wave applied at the input. The four important transient parameters for the inverter are the low-to-high propagation delay, t_{PLH}, the high-to-low propagation delay, t_{PHL}, the output rise time, t_R, and the output fall time, t_F.

The rise time and fall time for the output are measured between the 10 and 90% points on the waveform. The 10% point is the point in time at which the output voltage is 1/10 of the way from the low value to the high value; the 90% point is defined in similar fashion. The propagation delays are measured between the 50% points on the input and output waveforms (the points at which the voltage is midway between the two limiting values). As a matter of nomenclature, the low-to-high propagation delay, t_{PLH}, refers to the low-to-high transition at the *output node*. For an inverter, this corresponds to the opposite transition at the input node. Similarly, t_{PHL} refers to the high-to-low transition at the output node. Often the two propagation delays are very different and should be specified. In other situations, the transient response is entirely symmetric so that the two propagation delays are equal. In such cases it suffices to use the notation t_P.

An inverse relationship exists between the worst-case propagation delay and the maximum achievable system clock frequency. As a rule of thumb,

$$f_{CLK} < \frac{1}{20t_p}, \tag{1.14}$$

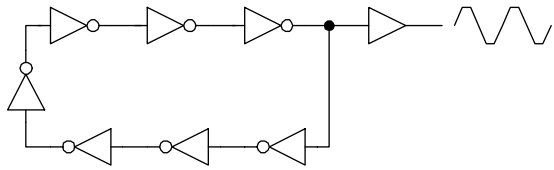


FIGURE 1.16
Seven-stage ring oscillator with a buffered output.

where f_{CLK} is the clock frequency. Of course, this rule is not hard and fast; the actual relationship depends on the efficiency of the system design. It is true, nonetheless, that an improvement in the propagation delay allows a direct improvement in the clock frequency and system performance.

Propagation delays are determined experimentally using ring oscillators such as the one shown in Figure 1.16. A ring oscillator comprises an odd number of inverters and the output at any one node is seen to oscillate with a frequency given by

$$f_M = \frac{1}{M\left(t_{PLH} + t_{PHL}\right)}, \tag{1.15}$$

where M is the (odd) number of inverters in the ring.

1.2.6 Power Delay Product

The power delay product (PDP) is an important figure of merit for a logic gate. It is defined by

$$\text{PDP} = Pt_p, \tag{1.16}$$

where P is the power dissipation per gate and t_p is the propagation delay. If t_{PLH} and t_{PHL} are different in value, then the average is used. The power dissipation may be a function of frequency, so the measurement conditions must be given in any meaningful specification of the PDP. The PDP has units of energy and typical values are measured in picojoules. Therefore, one interpretation is "the energy required to make a decision" and it is desirable to minimize the PDP.

The popular use of the PDP as a figure of merit reflects the power–speed trade-off in digital circuitry. For any given circuit, scaling the resistances will tend to provide an even trade-off between speed and power. Doubling the resistances will halve the power but double the propagation delays. Halving the resistances will double the power but halve the propagation delays. Either approach leaves the PDP unchanged; therefore, improvement of the power delay product requires improvement in the circuit or transistors. This is the essence of digital design at the device and circuit levels.

1.3 Logic Families

There are numerous families of logic gate circuits; they go by the acronyms NMOS, CMOS, BiCMOS, RTL, DTL, I^2L, TTL, ECL, and DCFL, to name a few. Over 100 families and subfamilies of logic gates exist. Each of these has unique properties that make it the most suitable for certain applications. Fortunately, this multitude of logic families can be broadly classified into four groups: MOS (metal oxide–semiconductor), saturated bipolar, nonsaturated bipolar, and compound semiconductor types.

MOS logic circuits are by far the most important today. They draw their name from the silicon MOSFETs on which they are based. The two important variations are NMOS, which uses n-channel MOSFETs exclusively, and CMOS, which uses p-channel and n-channel MOSFETs in complementary pairs. CMOS is used almost exclusively in modern microprocessors and application-specific integrated circuits (ASICs). The unique advantage of CMOS is the low standby power dissipation.

Saturated bipolar logic families include resistor–transistor logic (RTL), diode–transistor logic (DTL), and the many families of transistor–transistor logic (TTL). A common feature of these circuits is that they use silicon bipolar transistors as saturated switches. Their advantage is that bipolar transistors exhibit higher transconductance than MOSFETs. Therefore, although these circuits are inherently slower than CMOS, they are less susceptible to capacitive loading. The result is that saturated bipolar gates can outperform CMOS in applications involving large-load capacitances such as motherboards.

Nonsaturated bipolar logic families are collectively called emitter-coupled logic (ECL). ECL gates are the fastest silicon digital circuits with regard to off-chip propagation delays and outperform saturated bipolar circuits by a considerable margin. Consequently, they are important in high-performance computing and high data-rate digital communications. In addition, the possible development of ECL circuitry using SiGe alloys or compound semiconductors may further boost the performance of ECL.

Figure 1.17 illustrates the life cycle for families of SSI, MSI, and LSI logic circuits. Following introduction, a particular logic family experiences growth in sales, peaking at maturity; then, saturation and, eventually, decline occur. Despite the rapid advance of the state of the art, the logic family life cycle has historically extended over decades. Although standard TTL is in decline, the ALSTTL, ASTTL, and FAST families of transistor–transistor logic are just peaking in sales; 100k ECL and ABT BiCMOS are experiencing strong growth. Most of the newly introduced logic families are variations of CMOS.

For several reasons, logic families based on compound semiconductors have always been at a disadvantage relative to their silicon counterparts. First, silicon wafers are available in large sizes with a high degree of perfection and at low cost. Second, silicon processing technology is relatively advanced compared to that for compound semiconductors. Third, the manufacturing infrastructure (materials, equipment, service) favors silicon over

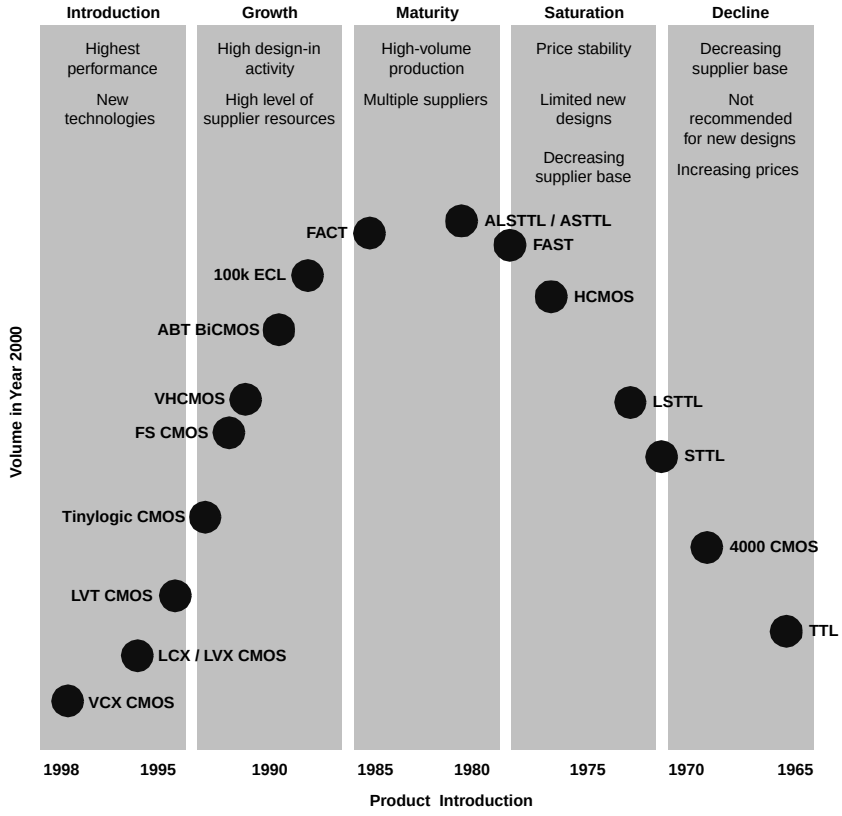


FIGURE 1.17
Logic family life cycle. The figure shows logic family sales in year 2000 vs. the year of product introduction. (Courtesy of Fairchild Semiconductor, South Portland, Maine.)

the lower-volume compound semiconductors. On the other hand, logic circuits based on compound semiconductors such as GaAs are capable of higher-speed operation than silicon circuits designed with similar layout rules. (GaAs has an advantage over Si by an approximate factor of three.) Therefore, compound semiconductor logic families such as GaAs direct-coupled FET logic (DCFL) are used only in niche applications that demand higher speed than is available from silicon. These include digital communication and, occasionally, microprocessors.

1.4 Computer-Aided Design and Verification

Modern digital integrated circuits have reached a degree of complexity that mandates the use of computer tools for design and verification. This is true

at every level of the design, from materials to devices to circuits to systems. As a result, computer tools are often vertically integrated to address all four levels of design. Increasingly, these computer tools use standardized languages and data files to ease design transfer between departments or corporations.

Design at the materials and device levels is closely inter-related and done together in practice. Sophisticated process simulation tools such as Silvaco and Cadence are used to arrive at the dimensions and process steps used for the fabrication of the core transistors. Any particular chip design will typically use fewer than four distinct transistor device designs. Therefore, these core transistor designs are used repeatedly in the circuit designs. Here computer tools are indispensable in the verification of circuit performance. SPICE (*simulation program with integrated circuit emphasis*) is a circuit simulation program developed at Stanford from which all modern circuit simulator programs are descended. These modern simulators go by a number of names including HSPICE and PSPICE.[20] Some are integrated in complete chip design packages, such as Cadence SPICE.

Design at the system level of abstraction begins with a library of logic gate circuits used repeatedly throughout the design. This streamlines the process in much the same way as using a device library for circuit design. The standard design languages used here are Verilog and VHDL, which stands for VHSIC (very high-speed integrated circuits) hardware description language.

The device, circuit, and process designs are inseparably linked together. Thus, the product of the design process is a set of computer-generated photomasks that are used for pattern transfer to the wafer during processing steps as described in the next section.

1.5 Fabrication

Modern digital integrated circuit fabrication involves a complex sequence of processing steps[21] that result in a packaged chip. This sequence begins with the growth of large cylindrical boules of doped single crystal silicon. Even boules of 100 to 200 kg are grown virtually free from crystal defects. Following crystal growth, boules are sawed into 200- to 300-mm diameter round wafers. These wafers are chem-mechanically polished to a mirror finish and a final thickness of 0.375 to 0.500 mm and subjected to a series of processing steps with a goal of creating all necessary devices and wiring them together in the desired fashion. The basic steps include epitaxial growth of single-crystal layers of silicon, doping by ion implantation or solid-state diffusion, oxidation, and deposition of various layers including insulators, polycrystalline silicon (polysilicon), and metals. With few exceptions, patterns are defined on the wafer at each step using photomasks and a process called photolithography.

Five basic process sequences are in use today: CMOS,[22–24] DRAM,[25] bipolar,[21,26,27] BiCMOS,[28,29] and GaAs E/D MESFET.[30,31]

- The CMOS (complementary metal oxide–semiconductor) process is designed to provide n-channel and p-channel MOSFETs on the same wafer. This process has many variations, but the generic process flow outlined in Section 1.5.1 captures the salient features of CMOS fabrication. A particularly important variation on this process is silicon on insulator (SOI).[32]
- The DRAM (dynamic random access memory) process is designed to provide n-channel MOSFETs and MOS storage capacitors on the same wafer. The basic steps in the n-channel MOSFET fabrication are similar to those used for the CMOS process. However, the MOS capacitor fabrication is done differently by every manufacturer so no attempt will be made to present a generic DRAM process here.
- The bipolar process is optimized for the fabrication of npn bipolar transistors and is used for ECL and TTL chips. Pnp devices can also be constructed, but with compromised performance. The process flow is outlined in Section 1.5.3.
- The BiCMOS (bipolar–CMOS) process is designed to provide npn bipolar transistors as well as CMOS devices.
- Finally, the GaAs E/D MESFET process is designed to provide enhancement type and depletion type metal–semiconductor field-effect transistors (MESFETs) on the same chip. It is used for GaAs digital integrated circuits such as direct-coupled FET logic (DCFL) circuits, which require normally off (enhancement type) and normally on (depletion type) MESFETs. The GaAs E/D MESFET process flow is outlined in Section 1.5.4.

Common to all of these processes is the need for multiple levels of metal for intrachip connections and metal pads for making interchip connections. Following transistor fabrication, a layer of insulator is deposited. Then windows are opened in this insulator, using photolithography, for contacting the devices. Next a layer of metal is deposited and patterned using photolithography. Subsequent layers of insulator and metal are put down and patterned in succession, with as many as 8 to 10 layers of metal used in VLSI circuits. Finally, large square metal pads are established for making electrical connections to other chips and devices. In some cases, it is adequate to arrange bonding pads around the periphery of the chip; in others, the pin count is such that metal pads must be arranged in a grid. This is true for modern microprocessors, which have hundreds of pins.

The following sections outline generic sequences for the CMOS, bipolar, and GaAs E/D MEFET processes and focus on details particular to the process at hand. Thus, many of the details of the photolithographic pattern

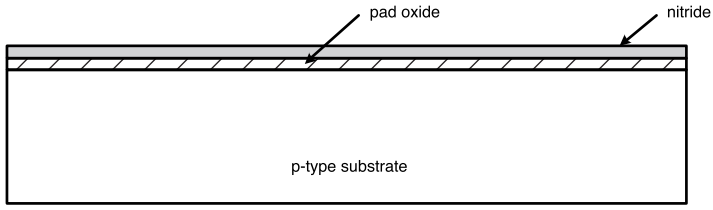


FIGURE 1.18
CMOS process. Pad oxide is grown and nitride is deposited.

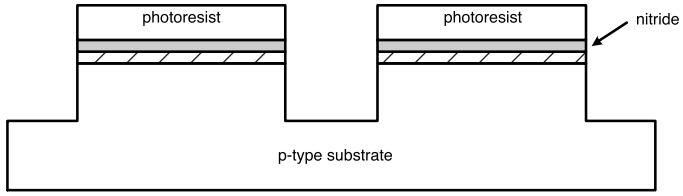


FIGURE 1.19
CMOS process. Trenches are etched using patterned photoresist.

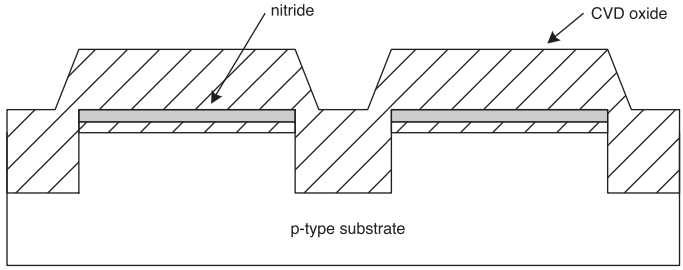


FIGURE 1.20
CMOS process. Oxide is deposited.

transfer have been omitted and the multiple levels of metal interconnect have not been shown.

1.5.1 CMOS Process

Although the CMOS process has many variations, they are all essentially similar to the generic sequence described next. The starting wafers are p-type silicon. Following oxidation and the deposition of silicon nitride as shown in Figure 1.18, trenches are etched. The pattern of trenches is defined by a photolithographic process using photoresist, after which reactive ion etching is used to process steep trench sidewalls as shown in Figure 1.19. The blanket deposition of silicon dioxide (Figure 1.20) by *chemical vapor deposition* (CVD) is followed by *chem-mechanical polishing* (CMP), which produces an optically flat surface with regions of embedded *shallow trench oxide* (STO) as shown in Figure 1.21.

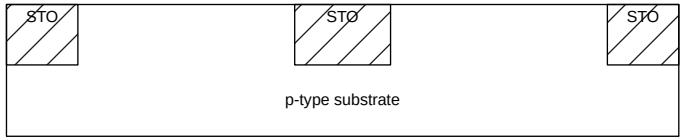


FIGURE 1.21
CMOS process. The surface is planarized by chem-mechanical polishing.

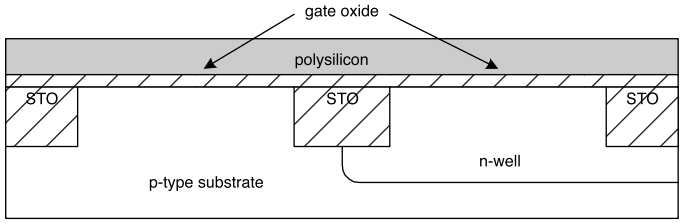


FIGURE 1.22
CMOS process. Gate oxide is grown and polysilicon is deposited.

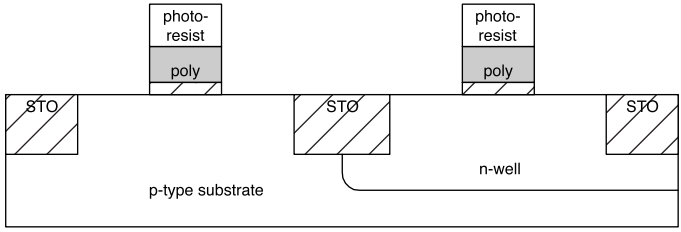


FIGURE 1.23
CMOS process. The gate oxide and polysilicon are patterned.

After the shallow trench definition, n-type wells are formed by ion implantation of phosphorus. Two additional ion implantation steps are used to fine-tune the threshold voltages of the MOSFETs. P-doping (boron) and n-doping (phosphorus) are used to adjust the n-channel and p-channel devices, respectively. Next a thin gate oxide is formed by dry thermal oxidation, followed by blanket chemical vapor deposition of polysilicon (see Figure 1.22). The polysilicon and gate oxide are patterned to form the MOS gate structure, as shown in Figure 1.23. (Despite their name, all modern MOSFETs use polysilicon, rather than metal, gates.) These gate structures act as masks for the n^+ (arsenic) and p^+ (boron) implantations used for the sources and drains of the n- and p-channel devices, respectively (Figure 1.24). This automatically aligns the source and drain edges with the gate edges, improving device performance. An added benefit is that the polysilicon gates are doped at the same time as the sources and drains.

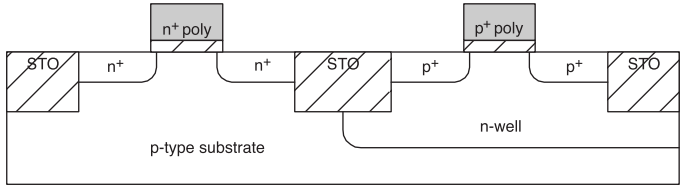


FIGURE 1.24
CMOS process. The source and drain regions are ion implanted using the gates as masks (self-aligned process).

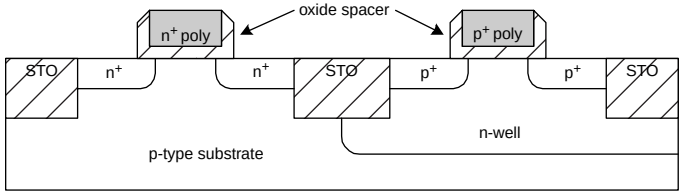


FIGURE 1.25
CMOS process. Spacers are produced using anisotropic etching.

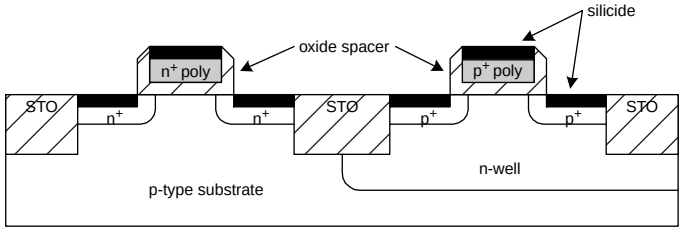


FIGURE 1.26
CMOS process. Silicide contacts are formed.

Next, oxide spacers are placed between the gate structures and the source/drain regions (Figure 1.25) by blanket CVD oxide and anisotropic (highly directional) etching. Then low-resistance silicide contacts are formed. Silicides are compounds formed between silicon and metals (Figure 1.26); for example, titanium silicide may be formed by the deposition of 200 Å of titanium, followed by reaction with the silicon at high temperature and etching of the unreacted titanium. The fabrication is completed by the deposition and patterning of alternating levels of metal interconnect and insulating material. As many as 10 levels of metal are used at the present time.

1.5.2 Silicon on Insulator (SOI)

Because silicon on insulator[32] is an important variation on the CMOS process described in the previous section, it warrants a separate description. Broadly

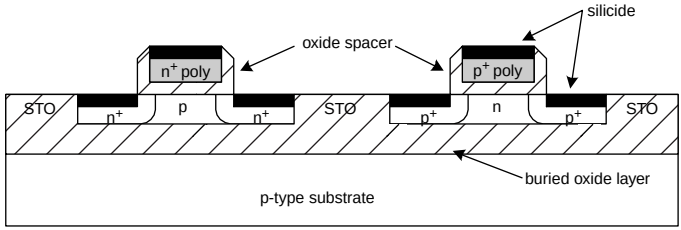


FIGURE 1.27
Silicon-on-insulator (SOI) CMOS devices fabricated using the separation by implantation of oxygen (SIMOX) process.

speaking, SOI refers to any realization of silicon integrated circuits on an insulator. The advantages of SOI are reduced parasitic capacitances, superior subthreshold characteristics in MOSFETs, and improved radiation hardness. Lower parasitic capacitances are achieved in SOI due to the avoidance of p–n junctions to isolate devices. Instead, each device is fabricated in a tub of silicon dioxide (Figure 1.27), which has a lower permittivity than silicon. The lower parasitics lead to modest improvements in propagation delays (~25% improvement). SOI MOSFETs have superior subthreshold characteristics compared to conventional CMOS devices. This, combined with the absence of the body effect in the MOSFETs, makes SOI especially suitable for low-power CMOS integrated circuits.

SOI circuitry is radiation hard compared to conventional silicon integrated circuits. Ionizing radiation incident on a silicon integrated circuit creates excess electrons and holes that affect the operation of the transistors in the circuit. This can lead to bit errors, also known as "soft errors." The generation of electron–hole pairs is a volume effect associated mostly with the bulk of the silicon substrate. The electrical isolation of the substrate from the circuits by an insulating layer greatly reduces the susceptibility of the circuit to such soft errors.

Particular approaches to SOI include silicon on sapphire (SOS), wafer bonding and thinning, and separation by implantation of oxygen (SIMOX). SOS involves depositing polycrystalline silicon on sapphire, followed by a recrystallization process. The drawbacks of SOS include the lack of large-diameter wafers and inability to produce single-crystal silicon of adequately high quality. Wafer bonding involves bonding a silicon wafer to an insulating wafer, followed by CMP, to leave only a thin layer of silicon over the entire wafer. This process is expensive and requires precise control in the CMP step. SIMOX involves the ion implantation of oxygen into a single crystal silicon wafer to create a buried layer of silicon dioxide. The depth of the silicon dioxide can be controlled precisely by the choice of the ion implantation energy. Other schemes for SOI have been developed as well; however, the SIMOX process has been very successful because it leverages the silicon wafers and ion implantation equipment already in use for bulk CMOS fabrication. This process results in the structure shown in Figure 1.27.

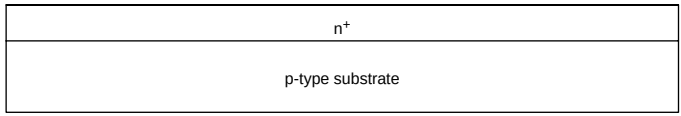


FIGURE 1.28
Bipolar process. n^+ epitaxy on p-type substrate (subcollector).

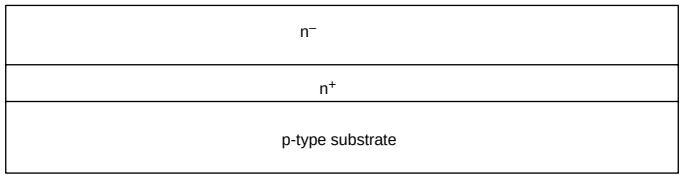


FIGURE 1.29
Bipolar process. Lightly doped n-type epitaxial layer (collector).

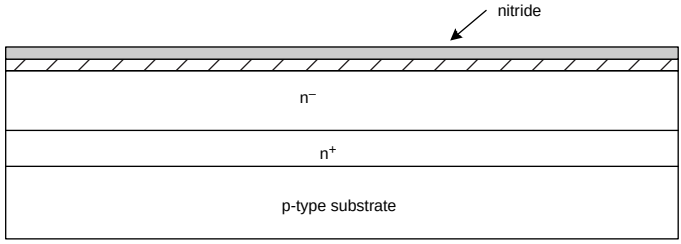


FIGURE 1.30
Bipolar process. Oxide growth and nitride deposition.

1.5.3 Bipolar Process

Early bipolar transistors were junction isolated and fabricated by the double-diffused epitaxial (DDE) process. This approach gave way to oxide isolation in the form of LOCOS (local oxidation of silicon oxide) and, finally, trench isolation. Trench-isolated devices are smaller in size and provide improved performance over the other types of bipolar transistors. Consequently, they are used in all high-performance bipolar digital circuits today. The following figures outline a modern process based on deep trench isolation.

The process begins with arsenic ion implantation of an n+ layer as shown in Figure 1.28. This layer is subsequently buried by epitaxy (growth of a single crystal layer of silicon from vapor phase sources) as shown in Figure 1.29. The wafer is then oxidized to create a layer of silicon dioxide and coated by CVD nitride as shown in Figure 1.30. Once patterned, the nitride is used as a mask for the etching of deep trenches by reactive ion etching techniques as shown in Figure 1.31. These deep trenches are oxidized and back-filled

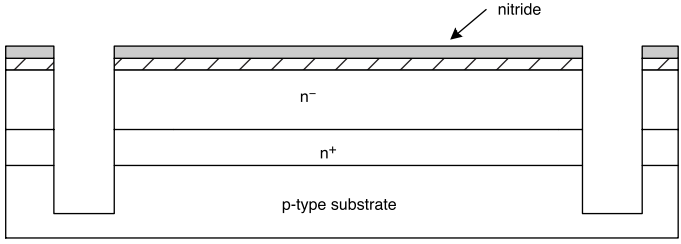


FIGURE 1.31
Bipolar process. Etching of deep trenches.

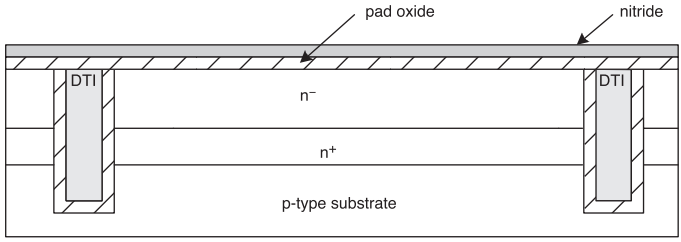


FIGURE 1.32
Bipolar process. Deep trenches are oxidized and backfilled with polysilicon. Oxide is grown and nitride is deposited.

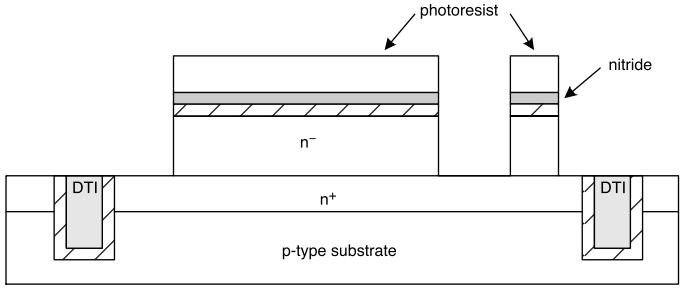


FIGURE 1.33
Bipolar process. The n-type epitaxial layer is patterned.

with CVD polysilicon. After CMP, the wafers are reoxidized and another nitride layer is deposited, resulting in the structure shown in Figure 1.32. Pattern definition (Figure 1.33) is followed by deposition of oxide (Figure 1.34) and CMP, producing shallow trench oxide (STO) as shown in Figure 1.35.

Boron-doped p-type polysilicon is deposited for the base contacts (Figure 1.36) and sidewall oxide spacers are formed by deposition and patterning (Figure 1.37). Then the extrinsic p+ base is diffused from the p+ polysilicon (Figure 1.38), followed by formation of a pedestal collector beneath the wafer

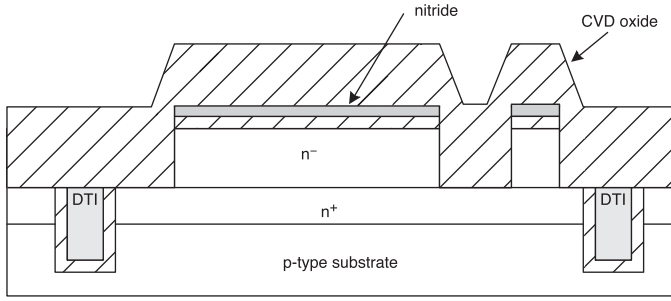


FIGURE 1.34
Bipolar process. Oxide is deposited by chemical vapor deposition for shallow trench oxide.

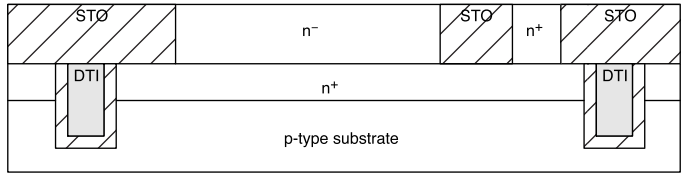


FIGURE 1.35
Bipolar process. The surface is planarized by chem-mechanical polishing (CMP).

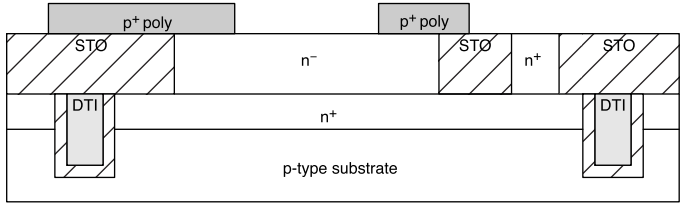


FIGURE 1.36
Bipolar process. The p^+ polysilicon base contacts are deposited and patterned.

surface by ion implantation of phosphorus (Figure 1.39). Next, the intrinsic (active) part of the base is doped by boron ion implantation (Figure 1.40). The device fabrication is completed by the deposition of an n+ polysilicon emitter (Figure 1.41) and the formation of silicide contacts to the emitter, base, and collector (Figure 1.42). Chip fabrication is completed with the addition of multilayer metal interconnects, bonding pads, and a protective cover layer.

1.5.4 GaAs E/D MESFET Process

The GaAs E/D MESFET process is unique in two respects. First, the material used is gallium arsenide, a compound semiconductor; second, the devices are metal–semiconductor field-effect transistors (MESFETs). The name of the

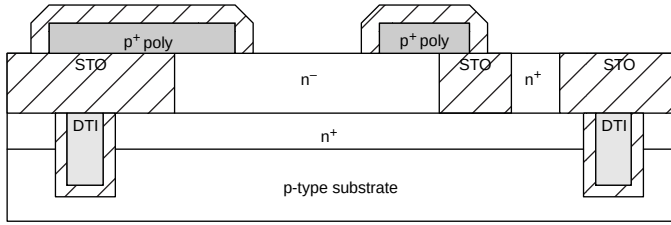


FIGURE 1.37
Bipolar process. Oxide is grown and patterned.

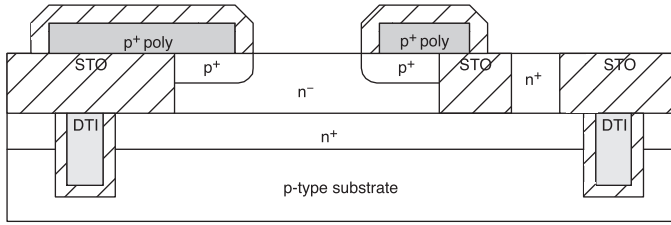


FIGURE 1.38
Bipolar process. The p-type base is diffused from the p^+ polysilicon.

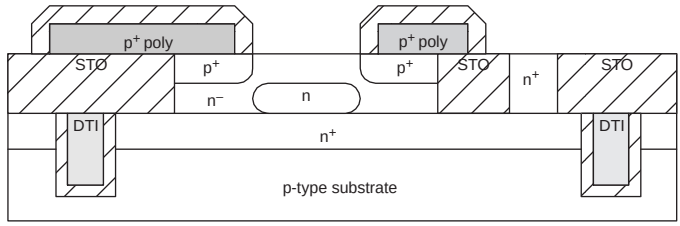


FIGURE 1.39
Bipolar process. The n-type collector is implanted.

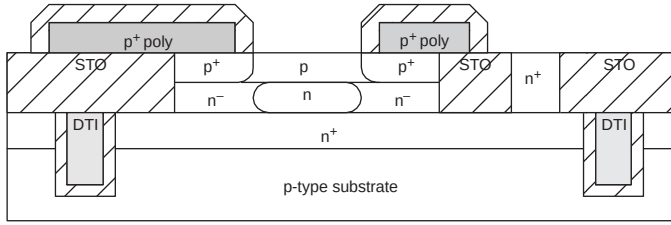


FIGURE 1.40
Bipolar process. The p-type base is implanted.

process derives from the fact that enhancement type transistors (E-MESFETs) and depletion type devices (D-MESFETs) are made on the same wafer. This is necessary for the commonly used GaAs digital logic families such as direct-coupled FET logic (DCFL).

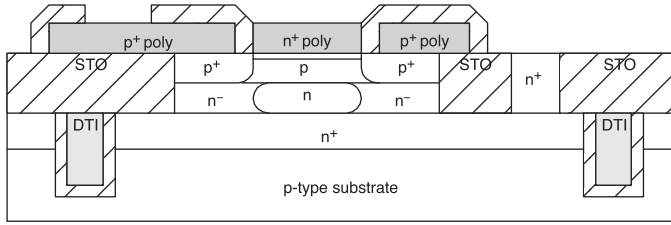


FIGURE 1.41
Bipolar process. The n^+ polysilicon emitter is formed.

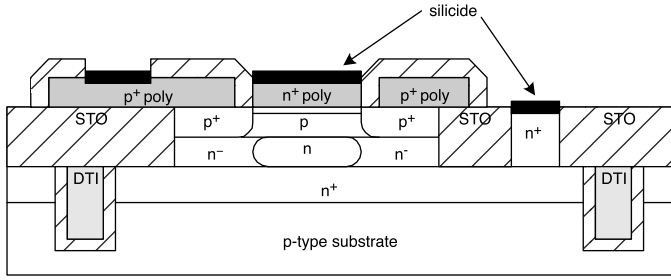


FIGURE 1.42
Bipolar process. Silicide contacts are formed.

At the present time, gallium arsenide wafers are expensive and small in size (100 vs. 300 mm) compared to silicon wafers. In its favor, electrons in this semiconductor have a higher low-field mobility resulting in shorter gate delays. Also, the electron velocity vs. electric field characteristic is such that GaAs circuits achieve lower power-delay products than their silicon counterparts. A final advantage of gallium arsenide wafers is that, unlike silicon, they are available in truly semi-insulating form (resistivity > 10^9 Ωcm). This makes the devices self-isolating and simplifies the overall fabrication process.

The E/D MESFET process begins with the deposition and patterning of a mask layer, which is then used to define the ion-implanted source and drain regions for MESFETs. Silicon or sulfur implants are used here (Figure 1.43). After this mask layer is stripped, another is deposited and patterned to define the n-type channel implantation for the depletion type device (Figure 1.44); the n-type channel implantation for the enhancement type device is done in a separate step (Figure 1.45) because the two types of devices must have different pinch-off voltages. Next, Ti/Pt/Au gate metal is deposited by electron beam evaporation and then patterned (Figure 1.46). The process is not self-aligned because of the need to form the source and drain regions before the gates. Therefore, an additional ion implantation step must be performed (Figure 1.47) to correct for any misalignment between the gates and the source–drain regions. Device fabrication is completed by the deposition and patterning of an insulator, the formation of alloyed Au/Ge ohmic contacts (Figure 1.48), and the opening of contact windows (Figure 1.49). As usual,

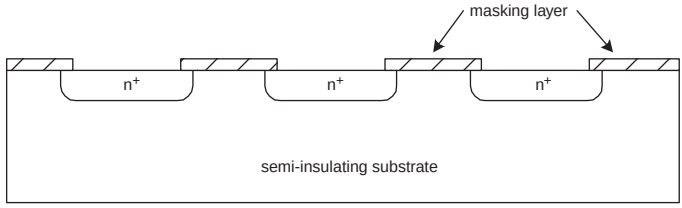


FIGURE 1.43
GaAs E/D MESFET process. Source and drain regions are implanted into the semi-insulating GaAs substrate.

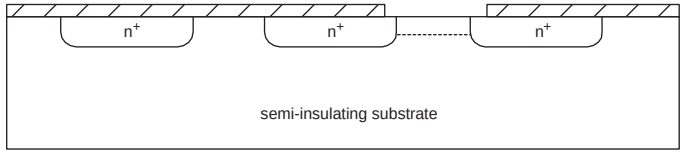


FIGURE 1.44
GaAs E/D MESFET process. The channel of the DFET is implanted.

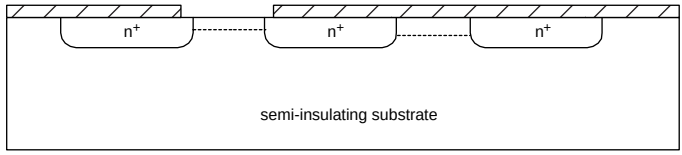


FIGURE 1.45
GaAs E/D MESFET process. The channel of the EFET is implanted.

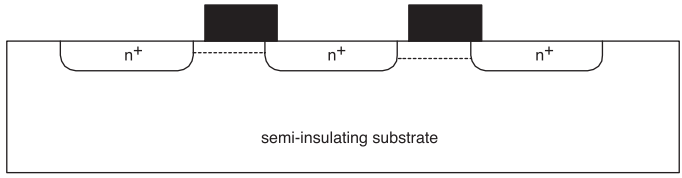


FIGURE 1.46
GaAs E/D MESFET process. The gate metal is deposited and patterned.

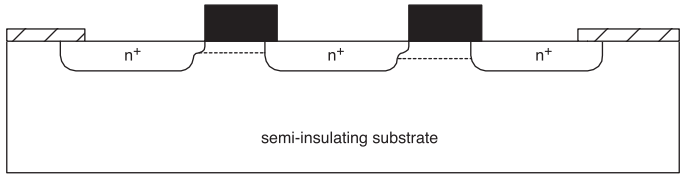


FIGURE 1.47
GaAs E/D MESFET process. An n^+ implant self-aligns the gates.

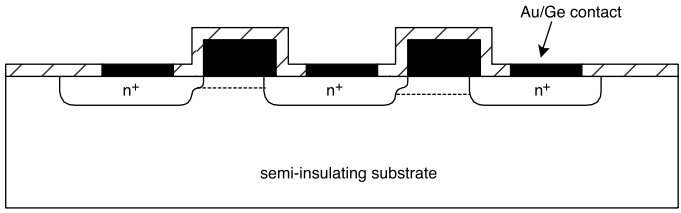


FIGURE 1.48
GaAs E/D MESFET process. Au/Ge ohmic contacts are formed.

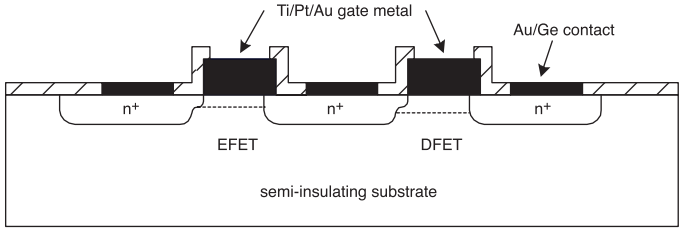


FIGURE 1.49
GaAs E/D MESFET process. Windows are opened for gate contacts and the circuit is ready for multiple layers of metallization.

circuit fabrication is completed with multiplayer metal interconnects, bonding pads, and a protective cover layer.

1.6 Testing and Yield

Testing is necessary to verify the functionality of fabricated integrated circuits. *Wafer testing* is always performed on the wafer prior to dicing and packaging; each circuit is contacted by electrical probes during testing. These wafer probes take the form of small needles that make electrical contact with the bonding pads at the periphery of the circuit. Obviously, accurate alignment is required for this probing. Once electrical contact is made, computer-generated test signals are applied to the circuit and the resulting outputs are measured to evaluate the performance of the circuit. If a malfunction is detected, a dot of ink is applied to the circuit to indicate that it is bad. This allows the manufacturer to avoid the expense of packaging bad die.

An important economic consideration is the yield, which is the fraction of good circuits on a wafer after fabrication. Although manufacturers seldom disclose their yields, values obtained in practice vary from about 80% up to more than 95%, depending on the nature of the circuit and the process. However, the yield critically affects the profit margin, so only yields exceeding 90% are economically viable for mainstream silicon CMOS products.

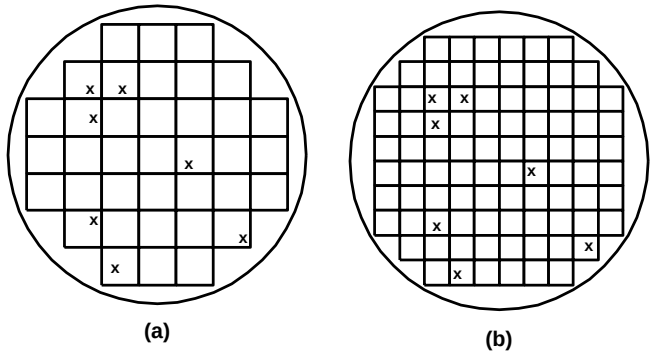


FIGURE 1.50
Effect of die size on yield. Both wafers have the same number and distribution of defects. For wafer (a) the yield is 81%, whereas for wafer (b) the yield is 92%.

The yield is a function of the wafer defect density and the die size. Assuming that a single defect results in a nonworking circuit, larger die* will result in a lower yield, even for the same defect density and distribution, as shown in Figure 1.50. The identically sized wafers have the same number and distribution of defects. The wafer on the left has 7 bad die out of a total of 37; the yield is 30/37, or 81%. The wafer on the right with a smaller die size also has 7 bad die, but the yield is 81/88, or 92%. A yield of 92% is economically viable, but 81% is probably not.

The underlying causes for bad die may be broadly classified as processing variations and point defects. Processing variations in doping, oxide thickness, metal thickness, or epitaxial layer thickness can result in nonworking devices or circuits that do not meet the required electrical specifications. Mask alignment tolerances are also important for large wafers; therefore, the process tolerances imposed by VLSI circuits are extremely tight. Point defects originate as particulates (dust) in the air of the processing facility and are of great importance. For example, a 3-μm dust particle that lands on the slice during processing may cause a break in a 1-μm metal wire. For this reason, all wafer fabrication processes are carried out in *clean rooms* with specially filtered air. Class 100 clean rooms have a maximum of 100 particles per cubic foot of air (~3500/m^3) greater than 0.5 μm, and a maximum of 10 particles per cubic foot greater than 0.5 μm.** Class 10 clean rooms are used for critical processing steps such as photolithography.

A number of models have been developed for the prediction of yield; the simplest one is based on the assumption of a uniform defect density. Suppose the number of chips on one wafer is N and the number of defects on the wafer is N_D. Let N_G be the number of good chips on the wafer. If a defect is

* The plural of die is dice. However, it has become standard practice in the industry to use "die" as the plural.

** The cleanliness requirements imposed by VLSI processing exceed those of hospital operating rooms, which are typically class 1000.

added randomly to the surface of the wafer, the probability of its ruining a good chip is N_G/N. Thus

$$dN_G = -\left(\frac{N_G}{N}\right)dN_D \tag{1.17}$$

and the fractional yield is therefore

$$Y = \frac{N_G}{N} = e^{-N_D/N} = e^{-D_0 A}, \tag{1.18}$$

where D_0 is the areal density of defects in cm^{-2} and A is the die area in cm^2. In practice, this model understates the yield for large area die. For this reason, more complex models have been developed using nonuniform defect distributions.

Integrated circuits of increasing complexity have made it difficult or impossible to test the functionality of some chips, such as microprocessors, completely. Circuits are increasingly designed for ease of testing. In addition, special test circuits are often built into the wafer for this purpose. Design for test (DFT) is a field of intense research and development today.

1.7 Packaging

Following fabrication, wafers are diced into rectangular chips (or die) that must be packaged. In a broad sense, packaging involves attaching the die to a substrate, making electrical connections to the die, and enclosing the package. A multitude of package types exists. Although some have been standardized by the Joint Electron Device Engineering Council (JEDEC), others are unique to a single product or product line.

Any integrated circuit package must:

- Support and protect the integrated circuit from mechanical shock
- Provide protection against the chemical environment, including moisture
- Conduct heat away from the integrated circuit and to an appropriate heat sink
- Be able to withstand a range of operating temperatures and repeated thermal cycling without failure
- Provide the necessary electrical connections to the integrated circuit, without undue degradation of the circuit speed

All of these requirements must be met in a package that is inexpensive and small in size.

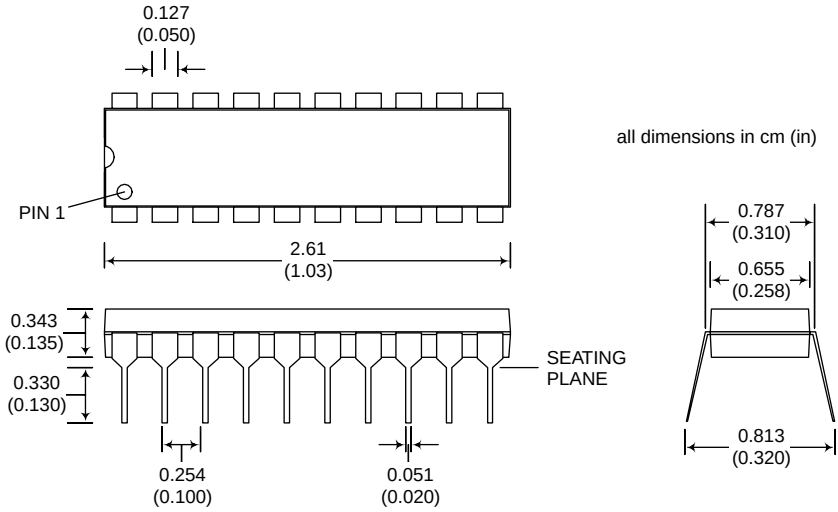


FIGURE 1.51
Plastic dual in-line package (PDIP) with 20 pins.

Broadly speaking, integrated circuit packages may be classified as through-hole packages, surface-mount packages, chip-scale packages, bare die, and module assemblies. Through-hole packages have metal pins that may be inserted through holes drilled in the circuit board for soldering. Surface-mount packages utilize metal leads that can be soldered to a single surface of the printed circuit board. Chip-scale packages are only slightly larger than the die they enclose and are attached to circuit boards via an array of solder bumps. Bare die are compact and avoid the electrical signal delays of a package, but are difficult to handle. Module assemblies combine several chips in one package.

Packages may be further classified as wire-bonded or flip-chip packages. Wire bonding involves using fine gold or aluminum wires to connect the bonding pads of the die to the package leads. The flip-chip approach involves mounting the chip face down; electrical connections to the package leads are made by solder bumps on the metal pads of the chip. A wire-bonded, through-hole type package is shown in Figure 1.51. This package is a type of dual in-line package (DIP) and has relatively few pins. Other package types can support more than 10^3 pins for VLSI circuits.

Chapter 18 will address these general principles in more detail. Some important modern package types will also be described in this context.

1.8 Reliability

Reliability is of utmost importance in all applications of digital integrated circuits.[33] Unlike other aspects of VLSIC performance, reliability is taken for

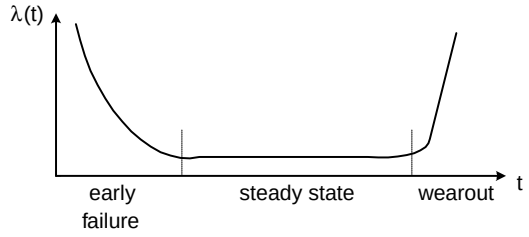


FIGURE 1.52
Typical failure rate vs. time for a component or system.

granted by customers. Most often, integrated circuit reliability is noticed by the customer *only if it is unsatisfactory*. Generally speaking, the probability $F(t)$ that an integrated circuit will fail at or before time t is a *cumulative distribution function* (cdf). This function approaches one as the time approaches infinity. The reliability function is the probability that the circuit will survive to time t without failure and is given by

$$R(t) = 1 - F(t) . \tag{1.19}$$

The *probability density function* (pdf) for circuit failure at time t is

$$f(t) = \frac{d}{dt} F(t) . \tag{1.20}$$

The instantaneous failure rate $\lambda(t)$ is

$$\lambda(t) = \lim_{\Delta \to 0} \left[\frac{1}{\Delta} \frac{R(t) - R(t+\Delta)}{R(t)} \right] = -\frac{d}{dt} \ln R(t) . \tag{1.21}$$

The mean time to failure (MTTF) is

$$MTTF = \int_0^\infty t f(t) dt . \tag{1.22}$$

In general, the failure rate for a component or system varies with time as shown in Figure 1.52. The initial high rate of failure is termed early failure or infant mortality and is due to manufacturing defects. The high failure rate near the end of life is due to wear-out. In the case of integrated circuits, wear-out may be eliminated almost entirely.

Various probability density functions have been used to model integrated circuit failure. The simplest is the exponential model, which results from a constant failure rate per device. In this case,

$$\lambda(t) = \lambda_0 = \text{ constant} , \tag{1.23}$$

TABLE 1.3

Effect of IC Failure Rate on System Reliability

IC failure rate (FIT)[a]	1	10	100
Failures per year (10^3 ICs per system)	0.009	0.088	0.88
% Systems failing in 5 years (10^3 ICs per system)	4.3	35.5	98.8

[a] 1 FIT = 1 failure in 10^9 device hours.

$$R(t) = e^{-\lambda_0 t}, \tag{1.24}$$

and

$$MTTF = \frac{1}{\lambda_0}. \tag{1.25}$$

Failure rates are measured in FITs (**F**ailure un**IT**s):

$$1\,\text{FIT} = \frac{1\,\text{failure}}{10^9\ \text{device hours}}. \tag{1.26}$$

For a system with 100 integrated circuits, a failure rate of 10 FIT corresponds to a mean time to failure of 114 years, whereas a failure rate of 1000 FIT corresponds to a mean time to failure of 1.1 year. A failure rate less than 10 FIT is very desirable, whereas 1000 FIT would be unacceptable in a complex system. Table 1.3 shows the impact of IC failure rate on the system lifetime, assuming the system contains 10^3 integrated circuits. It is important to note that early failure rates may be many times the steady-state failure rate.

The critical operating conditions that influence reliability are the supply voltage and temperature (junction temperature). Within the circuits, the most prevalent failure mechanisms are hot electron* injection (MOS circuits) and electromigration (MOS and bipolar circuits). The injection of hot electrons into the gate oxide of MOSFETs causes threshold voltage shift and device transconductance degradation, which eventually lead to circuit failure. Electromigration is an important failure mechanism in aluminum wires on the integrated circuit. Here, momentum transfer from the electrons to the aluminum atoms causes net transport of aluminum from areas of higher current density to areas of lower current density. This eventually leads to the open circuit failure of the interconnects. The most prevalent failure mechanisms at the package level are wire bond failure, solder fracture, and other mechanical fractures.

* Electrons in thermal equilibrium with the semiconductor lattice have an average energy of 3kT/2. Generally the term "hot electrons" applies to any situation in which the average electron energy exceeds this value. Such is the case in the high-field region near the drain of a short-channel MOSFET under normal operation.

Generally, all of the important integrated circuit failure mechanisms are thermally activated and can be fit by Arrhenius-type equations. Assuming that a particular failure mechanism is dominant, the lifetime of an integrated circuit can be modeled by

$$L = A \exp\left(\frac{E_a}{kT}\right), \tag{1.27}$$

where A is a constant, E_a is the activation energy, k is the Boltzmann constant, and T is the temperature in Kelvin. This simple model masks the dependence of the failure rate on the supply voltage. However, in practice the operating voltage is specified with a sufficiently tight tolerance to avoid this complication. Testing is therefore done under the worst-case condition (maximum value) for the operating voltage.

Operating temperature ranges for integrated circuits vary by application. Commercial-grade integrated circuits are intended for a temperature range of 0 to 125°C, whereas military (aerospace/defense)-grade integrated circuits are designed and tested for a wider temperature range of –50 to 125°C. Materials other than silicon must be used for higher temperature ranges. Silicon carbide and gallium nitride are examples of semiconductors that could be used to realize integrated circuits for high-temperature applications.

1.9 Burn-In and Accelerated Testing

Early failure, or "infant mortality," is an important reliability consideration for integrated circuits. Burn-in is useful to promote early failure before the integrated circuit is installed in a system.[34–43] During burn-in, the integrated circuits are exercised electrically while they are held at elevated temperature in an oven. The elevated temperature accelerates the failure mechanism or mechanisms responsible for early failure. However, the burn-in procedure must be designed in such a way that it does not compromise the reliability of the devices that survive it.[44]

As pointed out in the previous section, the typical failure rates for digital integrated circuits are so low that they may not be measured in a reasonable time scale. Here, accelerated testing may be used to estimate the actual failure rates. Most often, the acceleration is provided by an increased temperature. Then, assuming a thermally activated failure mechanism with an activation energy, E_a, the failure acceleration provided by an elevated temperature is given by

$$\text{acceleration} = \exp\left[\frac{E_a}{k}\left(\frac{1}{T_1} - \frac{1}{T_2}\right)\right]. \tag{1.28}$$

TABLE 1.4

Acceleration Factors for Temperature-Accelerated Testing of Integrated Circuits

	Acceleration Factor		Time Equivalent to 40 Years (h)	
T(°C)[a]	E_a = 1.0 eV	E_a = 0.5 eV	E_a = 1.0 eV	E_a = 0.5 eV
300	2.2×10^6	1500	0.2	233
250	3.2×10^5	570	1.1	616
200	3.1×10^4	176	11	2000
150	1700	41	200	2.0×10^4

[a] Ambient temperature = 60°C.

Table 1.4 gives acceleration factors for temperature-accelerated testing of integrated circuits. For a failure mechanism with an activation energy of 1.0 eV, testing at 150°C results in an acceleration factor of 1700 with respect to 60°C operation. Therefore, testing at 150°C for 200 h will produce as many failures as testing at 60°C for 40 years.

1.10 Staying Current in the Field

Staying current in the world's most dynamic field is a challenge. Success in this respect requires two components: a sound foundation of principles and diligent research. It is the purpose of this book to provide the former. Research in the field can best be undertaken using the myriad of online sources, a few of which are provided in Table 1.5. The International Technology Roadmap for Semiconductors characterizes the state of the art in all its details, with projections out to 15 years. Journal papers and conference proceedings provide the newest results obtained from industrial, academic, and government laboratories; industry trade journals provide information on companies and industrial practices and trends. Industry associations provide business statistics, sponsor conferences and meetings, influence government policies and regulations, and establish standards. Corporate sites, too numerous to tabulate here, provide valuable information on new and established products, including data sheets, application notes, and breaking news. In a field evolving so rapidly, the practices and players change appreciably from year to year, so online search engines are invaluable in finding new (or renamed) resources. Due to the very nature of the field, the information in this book should not be considered an end in itself, but rather a preparation for ongoing research on the topic.

1.11 Summary

Since the invention of the transistor in 1947 and the integrated circuit in 1958, progress in the field of integrated circuits has been extremely rapid. For

TABLE 1.5

Some Online Sources for Research in the Field

The International Technology Roadmap for Semiconductors	
public.itrs.net	International Technology Roadmap for Semiconductors
Journal papers, magazines, and conference proceedings	
ieeexplore.ieee.org	IEEE Explore
www.isinet.com	ISI Web of Science
Trade journals	
ap.pennet.com	Advanced Packaging
www.chipscalereview.com	Chip Scale Review
cr.pennwellnet.com	Cleanrooms
lucia.emeraldinsight.com	Microelectronics International
www.pennwell.com	Microlithography World
www.micromagazine.com	MICRO Magazine
www.semi.org	Semiconductor
www.e-insite.net/semiconductor	Semiconductor International
sst.pennwellnet.com	Solid State Technology
www.wafernews.com	WaferNews
Industry associations	
www.aeanet.org	American Electronics Association (AEA)
www.aset.or.jp	Association of Super-Advanced Electronics Technologies
www.edac.org	Electronic Design Automation Consortium (EDA)
www.eia.org	Electronic Industries Alliance (EIA)
www.mepteconline.com	Microelectronics Packaging and Test Engineering Council
www.semi.org	Semiconductor Equipment and Materials International (SEMI)
www.semichip.org	Semiconductor Industry Association (SIA)
www.svmg.org	Silicon Valley Manufacturing Group (SVMP)
Search engines	
www.altavista.com	AltaVista
www.aol.com	AOL
www.askjeeves.com	Ask Jeeves
www.ask.com	Earthlink
www.google.com	Google
www.lycos.com	Lycos
www.looksmart.com	Look Smart
www.msn.com	MSN
www.netscape.com	Netscape
www.go2.com	Overture (GoTo)
www.yahoo.com	Yahoo

several decades the maximum number of transistors per integrated chip has doubled every 18 months, as described by Moore's law. The International Technology Roadmap for Semiconductors (ITRS) maps out the expected future trends in integrated circuits with a 15-year horizon. The ITRS shows that known manufacturing solutions exist only for the next 5 years. Exponential progress and the need for rapid innovation will continue to make digital integrated circuits the most dynamic and challenging field of enterprise.

Basic properties of digital integrated circuits include logic function, fan-in and fan-out, voltage transfer characteristics, propagation delays, and power dissipation. Basic logic functions include NOT, OR, AND, NOR, NAND, and XOR. Fan-in refers to the number of circuit inputs, whereas the maximum fan-out, N_{MAX}, refers to the maximum number of similar load gates that can be connected to the output. The maximum fan-out may be limited by DC or by dynamic loading considerations. The voltage transfer characteristic is the output vs. input characteristic measured at low frequency. The propagation delay is the time required for a change at the input to affect the circuit output; it is measured using the 50% points on the input and output waveforms. The power dissipation includes DC and dynamic components. The power delay product (PDP) is an important figure of merit for a gate circuit.

Silicon logic families may be broadly classified as saturated bipolar, current mode bipolar, MOS, and BiCMOS. Saturated bipolar logic families include RTL, DTL, and TTL. Current mode bipolar logic gates include ECL. MOS logic families include NMOS and CMOS; BiCMOS circuits combine bipolar transistors with MOSFETs to achieve the advantages of MOS and bipolar logic. In addition to these, some logic families are based on semiconductors other than silicon, e.g., GaAs DCFL.

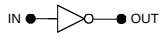
The fabrication of digital integrated circuits involves a sequence of many processes of doping, thin-film deposition, and etching. At each step, the required circuit patterns are transferred to the wafer using photolithography and photolithographic masks. Integrated circuit design involves the layout of these photomasks and is carried out using sophisticated computer tools. The basic fabrication processes in use are the CMOS, DRAM, bipolar, BiCMOS, and GaAs E/D MESFET processes.

Following fabrication, integrated circuits are tested on the wafer; bad circuits are marked with ink so that they may be discarded. The fraction of good chips (the yield) is an important economic consideration. Good die are packaged by a wire-bond or flip-chip process. Wire bonding involves the connection of fine gold wires between the chip and the package pins. The flip-chip approach involves turning the die face down and then making electrical connections using solder bumps on the integrated circuit. Packaged integrated circuits are tested prior to shipping. Because most integrated circuit failure mechanisms are thermally activated, accelerated testing is performed at elevated temperatures. Other failure mechanisms are induced by cyclic thermal stresses, so thermal cycle testing is also used.

DIGITAL INTEGRATED CIRCUITS QUICK REFERENCE

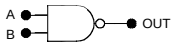
Moore's Law. The maximum number of transistors per chip doubles every eighteen months. The minimum feature size scales by 0.7 every eighteen months.

NOT



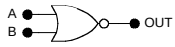
$Y = \overline{A}$

NAND



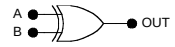
$Y = \overline{AB}$

NOR



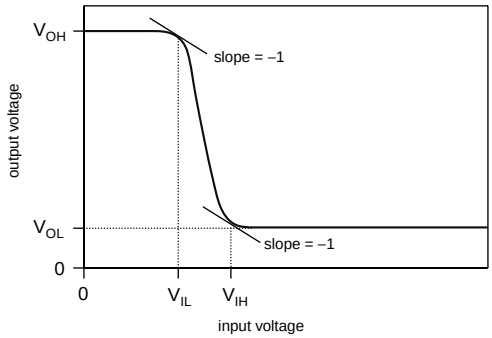
$Y = \overline{A + B}$

XOR



$Y = A \oplus B$

DC Voltage Transfer Characteristics



Fan-out

The maximum fan-out may be determined by DC or dynamic considerations:

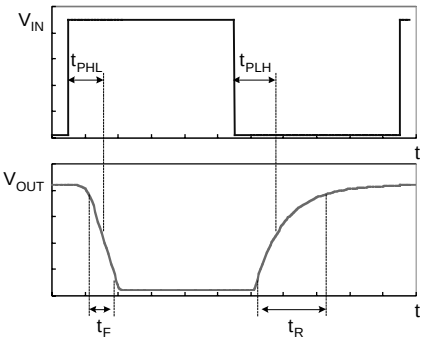
$$N_{MAX} < \frac{I_{OL}}{I_{IL}}; \quad N_{MAX} < \frac{I_{OH}}{I_{IH}}; \text{ and}$$

$$N_{MAX} < \frac{C_{L,\,MAX}}{C_{IN}}.$$

Logic Families

Bipolar (TTL, ECL), MOS (NMOS, CMOS), BiCMOS, GaAs DCFL

Dynamic Characteristics



Dissipation

$$P = P_{DC} + fC_L V^2_{DD}$$

$$P_{DC} = \frac{P_L + P_H}{2}$$

$$P_L = V_{DD} I_{DDL}$$

$$P_H = V_{DD} I_{DDH}$$

Power Delay Product

$$PDP = Pt_P$$

Fabrication Bipolar, CMOS, BiCMOS, DRAM, and GaAs E/D MESFET processes

Yield

$$Y = e^{-D_0 A}$$

Reliability

$$R(t) = 1 - F(t) \qquad f(t) = \frac{d}{dt} F(t) \qquad MTTF = \int_0^\infty t f(t)\,dt$$

Burn-in and Accelerated Testing

$$L = A\exp\left(\frac{E_a}{kT}\right) \qquad \text{acceleration} = \exp\left[\frac{E_a}{k}\left(\frac{1}{T_1} - \frac{1}{T_2}\right)\right]$$

f = 10^{-15} p = 10^{-12} n = 10^{-9} μ = 10^{-6} m = 10^{-3} k = 10^{3} M = 10^{6} G = 10^{9}

Problems

P1.1. Estimate the capacity for dynamic random access memories in the year 2030, assuming that industry will keep pace with Moore's law.

P1.2. As a consequence of Moore's law, the gate lengths in digital integrated circuits will eventually be scaled to the fundamental limitations imposed by silicon MOSFETs. Atoms in a silicon crystal are separated by 0.236 nm and MOSFET gate lengths certainly cannot be scaled to this level. When, according to the extrapolation of Moore's law, would gate lengths reach 1 nm? (Assume that gate lengths scale by 0.7 every 18 months.)

P1.3. The (uniform) areal density of defects for a CMOS fabrication process is 0.0063 mm^{-2}. Estimate the yield for 9×12-mm chips.

P1.4. The (uniform) areal density of defects for a CMOS fabrication process is 0.0014 mm^{-2}. Estimate the maximum economically viable chip size.

P1.5. When a type of integrated circuit is tested at 200°C, 10% of the circuits fail within 1000 h with V_{DD} = 1.65 V. Estimate the 10% lifetime at 175°C and 1.65 V, assuming E_a = 1.0 eV. Repeat for 100°C.

References

1. Bardeen, J. and Brattain, W.H., The transistor, a semiconductor triode, *Phys. Rev.*, 74, 230, 1948.
2. Early, J.M., Out to Murray Hill to play: an early history of transistors, *IEEE Trans. Electron. Devices*, 48, 2468, 2001.
3. Ross, I.M., The invention of the transistor, *Proc. IEEE*, 86, 7, 1998.
4. Brinkman, W.F., Haggan, D.E., and Troutman, W.W., A history of the invention of the transistor and where it will lead us, *IEEE J. Solid-State Circuits*, 32, 1858, 1997.
5. U.S. Patent 3,029,366 assigned to Sprague Electric Co., Massachusetts and U.S. Patent 3,025,589 assigned to the Fairchild Camera and Instrument Corp., New York.
6. Kilby, J.S., The integrated circuit's early history, *Proc. IEEE*, 88, 109, 2000.
7. Lilienfeld, J.E., U.S. Patent 1,745,175, 1930.
8. Kahng, D. and Atalla, M.M., Silicon–silicon dioxide field-induced surface devices, IRE Solid-State Device Res. Conf., Carnegie Institute of Technology, Pittsburgh, PA, 1960.
9. Arns, R.G., The other transistor: early history of the metal oxide–semiconductor field-effect transistor, *Eng. Sci. Educ. J.*, 7, 233, 1998.
10. Schaller, R.R., Moore's law: past, present and future, *IEEE Spectrum*, 34, 52, 1997.
11. Lucky, R.W., Moore's law redux, *IEEE Spectrum*, 35, 17, 1998.

12. Borkar, S., Obeying Moore's law beyond 0.18 μm (microprocessor design)], *Proc. 13th Annu. IEEE ASIC/SOC Conf.*, 26, 2000.
13. Sinha, A.K., Extending Moore's law through advances in semiconductor manufacturing equipment, *Proc. IEEE 1st Int. Symp. Qual. Electron. Design*, 243, 2000.
14. www.semichip.org (Semiconductor Industry Association Web site).
15. public.itrs.net (International Technology Roadmap for Semiconductors)
16. Allan, A., Edenfeld, D., Joyner, W.H., Jr., Kahng, A.B., Rodgers, M., and Zorian, Y., 2001 Technology roadmap for semiconductors, *Computer*, 35, 42, 2002.
17. Zeitzoff, P.M., Circuit, MOSFET, and front end process integration trends and challenges for the 180 nm and below technology generations: an international technology roadmap for semiconductors perspective, *Proc. 6th Int. Conf. Solid-State IC Technol.*, 1, 23, 2001.
18. Wakerly, J.F., *Digital Design: Principles and Practices,* 3rd ed., Prentice Hall, Upper Saddle River, NJ, 2002.
19. Hauser, J.R., Noise margin criteria for digital logic circuits, *IEEE Trans. Educ.*, 36, 363, 1993.
20. www.cadence.com (free download of the PSPICE student version).
21. Ghandhi, S.K., *VLSI Fabrication Principles*, 2nd ed., John Wiley & Sons, New York, 1994.
22. Davari, B., Chang, W.H., Wordeman, M.R., Oh, C.S., Taur, Y., Petrillo, K.E., Moy, D., Bucchignano, J.J., Ng, H.Y., Rosenfield, M.G., Hohn, F.J., and Rodriguez, M.D., A high performance 0.25-μm CMOS technology, *Tech. Dig. Int. Electron. Devices Meet.*, 56, 1988.
23. Luo, M.S.C., Tsui, P.V.G., Chen, W.-M., Gilbert, P.V., Maiti, B., Sitaram, A.R., and Sun, S.-W., A 0.25-μm CMOS technology with 45 Å NO-nitrided oxide, *Tech. Dig. Int. Electron. Devices Meet.*, 10, 691, 1995.
24. Baker, R.J., Li, H.W., and Boyce, D.E., *CMOS: Circuit Design, Layout, and Simulation*, IEEE Press, New York, 1998.
25. Kang, H.K., Kim, K.H., Shin, Y.G., Park, I.S., Ko, K.M., Kim, C.G., Oh, K.Y., Kim, S.E., Hong, C.G., Kwon, K.W., Yoo, J.Y., Kim, Y.G., Lee, C.G., Paick, W.S., Suh, D.I., Park, C.J., Lee, S.I., Ahn, S.T., Hwang, C.G., and Lee, M.Y., Highly manufacturable process technology for reliable 256-Mbit and 1-Gbit DRAMs, *Tech. Dig. Int. Electron. Devices Meet.*, 635, 1994.
26. Warnock, J.D., Silicon bipolar device structures for digital applications: technology trends and future directions, *IEEE Trans. Electron. Devices*, 42, 377, 1995.
27. Taur, Y. and Ning, T.H., *Fundamentals of Modern VLSI Devices*, Cambridge University Press, New York, 1998.
28. Kinoshita, Y., Suzuki, H., Nakamura, S., Fukaishi, M., Tajima, A., Sucmura, Y., Itani, T., Miyamoto, H., Fujii, H., Yotsuyanagi, M., Henmi, N., and Yamazaki, T., An advanced 0.25-μm BiCMOS process integration technology for multi-GHz communication LSIs, *Proc. Bipolar/BiCMOS Circuits Technol. Meet.*, 72, 1997.
29. Harame, D., High performance BiCMOS process integration: trends, issues, and future directions, *Proc. Bipolar/BiCMOS Circuits Technol. Meet.*, 36, 1997.
30. Long, S.I. and Butner, S.E., *Gallium Arsenide Digital Integrated Circuit Design*, McGraw-Hill, New York, 1990.
31. Mikkelson, J., GaAs digital VLSI device and circuit technology, *Tech. Dig. Int. Electron. Devices Meet.*, 231, 1991.
32. Adan, A.O., Naka, T., Kagisawa, A., and Shimizu, H., SOI as a mainstream IC technology, *Proc. SOI Conf.*, 9, 1998.

33. Hu, C., IC reliability simulation, *Proc. Int. Custom Integrated Circuits Conf.*, 1991.
34. Singh, A. and Rosin, J., Random defect limited yield using a deterministic model, *IEEE/SEMI Adv. Semicond. Manuf. Conf.*, 161, 2001.
35. Aloni, C., TYM system: an integrated tool for inherent line yield improvements for entire fab, *Proc. 9th Int. Symp. Semicond. Manuf.*, 237, 2000.
36. Ciplickas, D.J., Li, X., and Strojwas, A.J., Predictive yield modeling of VLSICs, *5th Int. Workshop Stat. Metrol.*, 28, 2000.
37. Maynard, D., Bombardier, S., Cavanaugh, A., and Zwonik, R., Modeling and optimization of wafer radial yield, *IEEE/SEMI Adv. Semicond. Manuf. Conf. Workshop*, 71, 1999.
38. El-Kareh, B., Ghatalia, A., and Satya, A.V.S., Yield management in microelectronic manufacturing, *Proc. 45th Electron. Components Technol. Conf.*, 58, 1995.
39. Drum, C.M., Applications of a mechanistic yield model for MOSIC chips, *Proc. Int. Workshop Defect Fault Tolerance VLSI Syst.*, 60, 1991.
40. Vollertsen, R.-P., Burn-In, *IEEE Int. Integrated Reliability Workshop Final Rep.*, 167, 1999.
41. Conti, D.R. and Van Horn, J., Wafer level burn-in, *Proc. 50th Electron. Components Technol. Conf.*, 815, 2000.
42. Vasco, F. and Lo, K., Relative effectiveness of thermal cycling vs. burn-in: a case study, Proc. *42nd Electron. Components Technol. Conf.*, 185, 1992.
43. Huston, H.H., Wood, M.H., and DePalma, V.M., Burn-in effectiveness-theory and measurement, *29th Annu. Proc. Reliability Phys. Symp.*, 271, 1991.
44. Taeho K., Kuo, W., and Chien, W.K, Burn-in effect on yield, *IEEE Trans. Electron. Packaging Manuf.*, 23, 293, 2000.

2

Semiconductor Materials

2.1 Introduction

The usefulness of semiconductor materials stems from the capability to control their conductivity by introducing small amounts of impurities. Materials may be broadly classified as insulators, semiconductors, semimetals, and metals. In an insulator like NaCl or SiO_2, all the valence electrons of the constituent atoms take part in strong ionic bonding and no free electrons are available for conduction of electricity. Metals, on the other hand, are typically made up of monovalent atoms like copper, gold or silver. All electrons are free to move around the solid and conduct electricity, resulting in high conductivity. In semiconductors and semimetals, all valence electrons take part in covalent bonding, so the intrinsic conductivity is low. However, some of these bonds can be broken by thermal energy, illumination by light, or introduction of impurities (doping); this property makes semiconductor materials useful in devices.

2.2 Crystal Structure

All materials used in digital integrated circuits are single crystals of the cubic class. These include silicon, which crystallizes with the diamond structure, and gallium arsenide, which crystallizes in the zinc blende structure. Some other interesting semiconductors, such as silicon carbide and gallium nitride, exhibit hexagonal structures but these will not be considered here.

Crystals are periodic arrangements of atoms in space. A crystal structure comprises a space lattice and a basis. The space lattice describes the periodic arrangement of points on which atoms or groups of atoms may be placed. The basis can be a single atom or an arrangement of atoms placed at each space lattice point. Of the 14 space lattices,[1] or Bravais lattices, 3 are cubic: simple cubic (SC), body-centered cubic (BCC), and face-centered cubic (FCC) space lattices. Figure 2.1 shows the conventional unit cell for the simple cubic

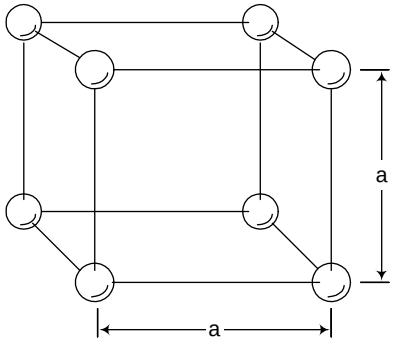


FIGURE 2.1
Simple cubic (SC) crystal structure.

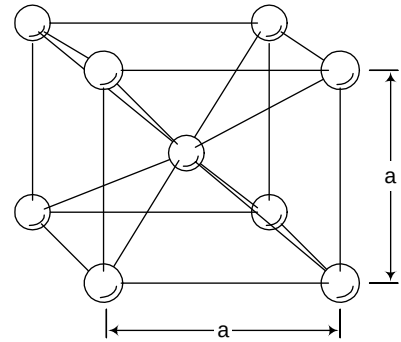


FIGURE 2.2
Body-centered cubic (BCC) crystal structure.

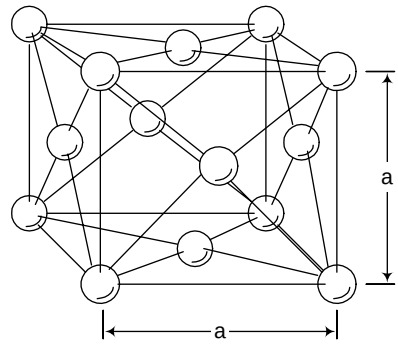


FIGURE 2.3
Face-centered cubic (FCC) crystal structure.

space lattice, which has a lattice point at each cube corner. The BCC conventional unit cell includes an additional lattice point at the cube center (Figure 2.2), while the FCC conventional unit cell has lattice points at the corners of the cube and also at the centers of each of the faces (Figure 2.3).

Silicon occurs in the diamond crystal structure, which can be constructed from an FCC space lattice with a basis of two atoms (Figure 2.4). The first basis atom is located on the space lattice point and the second is displaced from the first by one quarter of the cube diagonal. Thus, the diamond structure can be thought of as two interpenetrating FCC sublattices displaced from one another by one quarter of the cubic diagonal. Four atoms from the second FCC sublattice fall within the cubic unit cell of the first FCC sublattice. It should be recognized that the atoms on both sublattices are identical in the case of silicon. In gallium arsenide, however, one sublattice is made up entirely of gallium atoms and the second is made up entirely of arsenic atoms. In the silicon crystal, the lattice constant "a" (the side of the cube in the diamond lattice) is 5.4310 Å, whereas the lattice constant in GaAs is 5.6535 Å.

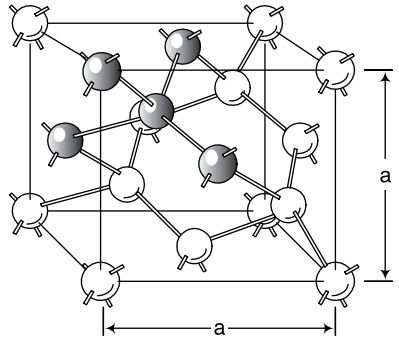


FIGURE 2.4
Diamond crystal structure. Each atom is tetrahedrally bonded to its four nearest neighbors (as shown by the five shaded atoms).

Atoms in a diamond or zinc blende crystal form tetrahedral bonds with the four nearest neighbors. In silicon the bonding is entirely covalent; thus a silicon atom with valence four shares electrons with four neighboring silicon atoms, resulting in a filled outer shell configuration. In gallium arsenide the bonding is primarily covalent but is partially ionic in nature due to the different numbers of valence electrons in gallium and arsenic.

The periodicity of the crystal lattice in real space also results in a periodicity of the potential energy experienced by electrons in the crystal. In turn, this gives rise to the energy band structure that dictates the electrical properties of a semiconductor.

2.3 Energy Bands

The unique properties of semiconductors result from the fact that the allowed energy levels of electrons exist in bands separated by forbidden gaps.[2–6] This behavior is due to the periodic crystalline structure and is quantum mechanical in nature. For individual atoms making up the crystal, discrete energy levels are allowed for electrons. If the atoms are brought together in close arrangement to form the crystal, their wave functions overlap and the discrete energy levels spread into bands in accordance with the Pauli exclusion principle. Figure 2.5 shows this general behavior schematically for a diamond in which the crystal contains N atoms per cubic centimeter. The isolated atoms have 2N "s" states and 6N "p" states; bringing the atoms together to form the crystal causes the "s" and "p" states to spread into bands. Further reduction of the atomic separation causes the bands to join and then split again. The result is two bands, each containing 4N states, separated by an energy gap. The crystal contains 4N valence electrons, so the valence (lower)

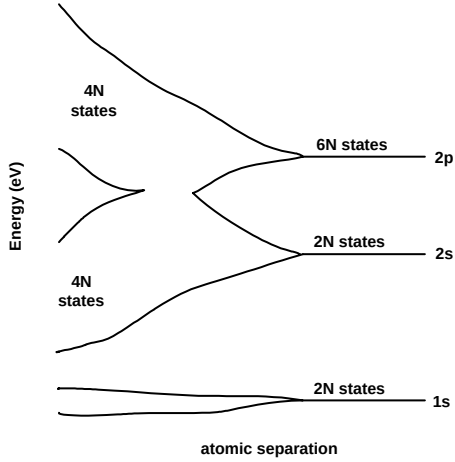


FIGURE 2.5
The formation of energy bands in diamond as isolated carbon atoms brought together to form a crystal. Silicon is qualitatively similar.

band is completely filled with electrons while the upper (conduction) band is completely empty at a temperature of 0 K.

Figure 2.6 shows the detailed energy band structure for silicon in an E vs. k diagram. Here, E is the electron energy and k is the electron wave vector. It can be seen that the crystal exhibits a number of energy bands, rather than discrete energy levels. The uppermost bands are conduction bands; mobile electrons reside near the bottom of the lowest conduction band. The lower bands are valence bands and mobile holes reside in the top valence band. The valence band and conduction band exhibit different curvature, resulting in different mobilities for electrons as holes, which will be explained in Section 2.6. The separation between the top of the valence band and the bottom of the conduction band is called the energy gap.

For most purposes, it is adequate to use simplified energy band diagrams of the type illustrated in Figure 2.7. Here, E_C is the bottom of the conduction band and E_V is the top of the valence band. E_f is the *Fermi level*,* which is a function of the doping concentration and temperature. E_i is the *intrinsic Fermi level*; it represents the position of the Fermi level in intrinsic (undoped) material. In n-type material, the Fermi level lies above E_i but in p-type material the Fermi level lies below E_i. The energy gap is the separation between the band edges, $E_C - E_V$. Simplified band diagrams of this type are useful for describing devices without undue complication. In a device that has a variation in the doping along one direction, the band diagram vs. distance may be represented without invoking a three-dimensional plot because of the omission of the wave vector dependence.

* The Fermi level represents the energy level at which a state has a 50% probability of being occupied by an electron. States below the Fermi level are more than half filled with electrons, whereas states above the Fermi level are less than half filled with electrons.

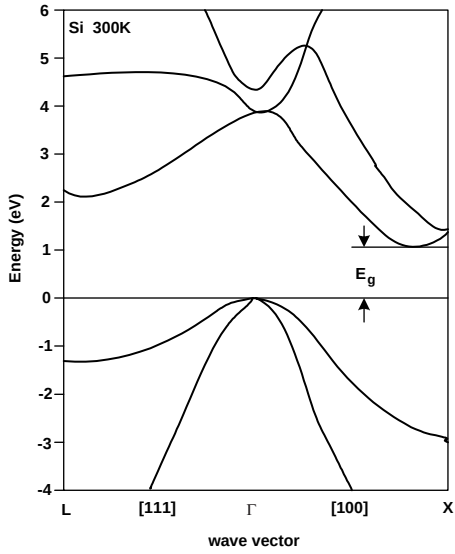


FIGURE 2.6
Detailed energy band structure of silicon at room temperature (300 K). The upper bands are the conduction bands and the lower bands are the valence bands. The energy gap E_g is the difference between the conduction band minimum and the valence band maximum.

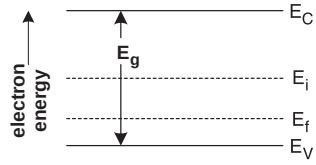


FIGURE 2.7
Simplified energy band diagram.

2.4 Carrier Concentrations

With all electrons bonded covalently, there are no free electrons to participate in electrical conduction unless some of the bonds are broken thermally, by illumination with light, or by doping. Regardless of how bond breaking is accomplished, the important result is the concentrations of free carriers (electrons and holes) that can carry electrical current. Free electrons exist in the conduction band with a concentration "n." The valence band is normally filled with electrons, so it is the absence of electrons that gives rise to conduction in this band. This behavior is described by considering the motion of fictitious positive charge carriers called holes, rather than the motion of

the remaining electrons in the band. (Holes in a valence band are analogous to bubbles in a beverage.)

In thermal equilibrium, the carrier concentrations depend on the position of the Fermi level relative to the band edges. The electron concentration is given by[3,5]

$$n = N_C e^{-(E_C - E_f)/kT}, \tag{2.1}$$

where N_C is the "effective density of states at the edge of the conduction band," E_f is the Fermi level, E_C is the edge of the conduction band, k is the Boltzmann constant, and T is the temperature in Kelvins. The hole concentration is given by

$$p = N_V e^{-(E_f - E_V)/kT}, \tag{2.2}$$

where N_V is the "effective density of states at the edge of the valence band" and E_V is the energy at the top of the valence band.

In intrinsic (undoped) silicon, the electron and hole concentrations are equal and E_f assumes the intrinsic Fermi level position, E_i. In this case,

$$n = N_C e^{-(E_C - E_f)/kT} = N_V e^{-(E_f - E_V)/kT} = p, \tag{2.3}$$

so that

$$E_f = E_i = \frac{E_C + E_V}{2} - \frac{kT}{2} \ln\left(\frac{N_C}{N_V}\right). \tag{2.4}$$

In intrinsic silicon, both carrier concentrations are equal to the intrinsic value n_i, given by

$$n_i = \sqrt{N_C N_V}\, e^{-E_g/2kT}, \tag{2.5}$$

where E_g is the bandgap. At room temperature, E_i is very close to the midgap and the intrinsic carrier concentration is about 1.4×10^{10} cm^{-3}.

The conductivity of a piece of semiconductor can be altered drastically by doping, which is the controlled introduction of impurities in small concentrations (usually parts per million). Arsenic, phosphorus, and antimony are pentavalent, so these atoms have five valence electrons. If one of these atoms replaces silicon in the crystal, four of the valence electrons take part in covalent bonding but the fifth electron is weakly bound to the phosphorus atom. Therefore, this fifth electron can readily break away at room temperature and be donated to the crystal as a free electron. For this reason, arsenic, phosphorus, and antimony are called donors. Boron, on the other hand, is trivalent and acts as an acceptor because, when boron replaces silicon in the crystal, one electron is missing in the covalent filled-shell configuration. The boron atom will readily accept an electron from the crystal, leaving behind a free hole.

Silicon doped with donors is called n-type. In this material, the donor ions are positively charged and the electrons are negatively charged. Therefore, by charge neutrality,

$$\overline{n} \approx N_d \,, \tag{2.6}$$

where $\overline{n}$ is the equilibrium electron concentration and N_d is the donor concentration, both in cm^{-3}. In equilibrium, the law of mass action applies so that

$$\overline{n}\,\overline{p} = n_i^2 \,; \tag{2.7}$$

therefore, for n-type material,

$$\overline{p} = \frac{n_i^2}{N_d} \,. \tag{2.8}$$

In n-type material the donor concentration is usually at least six orders of magnitude greater than n_i. Therefore, the concentration of electrons (majority carriers) is many orders of magnitude larger than the concentration of holes (minority carriers).

Silicon doped with acceptors is called p-type; in this material, the acceptors become negatively ionized. The carrier concentrations in p-type material are given by

$$\overline{p} \approx N_a \tag{2.9}$$

and

$$\overline{n} = \frac{n_i^2}{N_a} \,. \tag{2.10}$$

In p-type material, holes are the majority carriers and electrons are the minority carriers.

Example 2.1

Determine the carrier concentrations and the position of the Fermi level in a sample of silicon doped with phosphorus to a concentration of 10^{16} cm^{-3} and at a temperature of 300 K.

Solution. Phosphorus is a donor. The electron concentration is $n \approx N_d = 10^{16}$ cm^{-3} and the hole concentration is

$$p = \frac{n_i^2}{n} = \frac{\left(1.45 \times 10^{10}\ \text{cm}^{-3}\right)^2}{10^{16}\ \text{cm}^{-3}} = 2.1 \times 10^4\ \text{cm}^{-3} \,.$$

The position of the Fermi level relative to the intrinsic Fermi level is

$$\left(E_f - E_i\right) = \frac{kT}{q}\ln\left(\frac{n}{n_i}\right) = 0.0259\text{ eV}\ln\left(\frac{10^{16}\text{ cm}^{-3}}{1.45\times10^{10}\text{ cm}^{-3}}\right) = 0.348\text{ eV}.$$

The majority carrier (electron) concentration is many orders of magnitude greater than the minority carrier (hole) concentration and the Fermi level is much closer to the conduction band edge than the valence band edge.

2.5 Lifetime

In thermal equilibrium, the minority carrier concentration is dictated by the law of mass action, but may be much greater under nonequilibrium conditions. A departure from thermal equilibrium may be brought about by applying a voltage or irradiating with light, or by other means. In any case, the dynamics of the build-up and decay of the minority carrier concentration may be described by a time constant called the minority carrier lifetime (sometimes called simply "lifetime").[5]

Consider an n-type semiconductor illuminated by a lamp. If the photon energy is greater than the bandgap of the semiconductor, then each absorbed photon will break one bond. This will create one excess electron and one excess hole, so one electron–hole pair is created. The semiconductor remains space charge neutral, however. Thus

$$n' = p', \tag{2.11}$$

where n' and p' are the excess electron and hole concentrations, respectively, so that the total concentrations are given by

$$n = \overline{n} + n' \tag{2.12}$$

and

$$p = \overline{p} + p'. \tag{2.13}$$

Electrons and holes may recombine, annihilating each other and giving off an amount of energy equal to the bandgap. The recombination rate is proportional to the encounter rate and therefore the product of their concentrations. Thus,

$$R = Cnp = C\left(\overline{n} + n'\right)\left(\overline{p} + p'\right), \tag{2.14}$$

where R is the recombination rate and C is a constant. In n-type material with a moderate level of minority carrier injection, n' can be neglected compared to $\overline{n}$, so

$$R \approx C\overline{n}\left(\overline{p} + p'\right). \tag{2.15}$$

Therefore, the recombination rate for excess carriers is

$$R(n', p') = R(n, p) - R(\overline{n}, \overline{p}) = C\overline{n}p', \tag{2.16}$$

showing that the recombination rate for excess carriers is proportional to the excess minority carrier concentration.

Now suppose that the semiconductor is illuminated uniformly with a light, resulting in the generation of G_0 electron–hole pairs per second per cubic centimeter. Once the light is turned on, the net rate of change for the excess minority carrier concentration is the difference between the generation rate and the recombination rate. Thus

$$\frac{dp'}{dt} = G_0 - C\overline{n}p'. \tag{2.17}$$

If the initial excess minority carrier concentration is zero, then the solution is

$$p' = \frac{G_0}{C\overline{n}}\left(1 - e^{-tC\overline{n}}\right) = G_0\tau_p\left(1 - e^{-t/\tau_p}\right), \tag{2.18}$$

where τ_p is the minority carrier lifetime.

Now suppose the light has been on for some time and then is switched off at $t = 0$. With the light off,

$$\frac{dp'}{dt} = -C\overline{n}p' = -\frac{p'}{\tau_p}. \tag{2.19}$$

The initial condition is

$$p'(0) = G_0\tau_p \tag{2.20}$$

and the solution is thus

$$p' = G_0\tau_p e^{-t/\tau_p}. \tag{2.21}$$

An important conclusion of this analysis is that the minority carrier lifetime is the time constant for the build-up and the decay of the excess minority carrier concentration. In fact, the minority carrier lifetime can be defined on the basis of the decay rate:

$$\tau_p = \frac{p'}{-\dfrac{dp'}{dt}} = \frac{\left(\text{excess minority carrier concentration}\right)}{\left(\begin{array}{c}\text{time rate of decrease of}\\ \text{excess minority carrier concentration}\end{array}\right)}. \tag{2.22}$$

A similar analysis can be applied to a p-type semiconductor in which the minority carrier lifetime is τ_n.

The minority carrier lifetime in high-purity silicon is on the order of milliseconds. In high-speed silicon bipolar circuits, the minority carrier lifetime is intentionally shortened by doping with gold, platinum, or palladium, the so-called "lifetime killers." Typical device designs call for minority carrier lifetimes less than 1 ns, which require gold doping of about 10^{16} cm^{-3}.

2.6 Current Transport

Because they are mobile charge carriers, electrons and holes can transport electrical current. Underlying this current transport are drift, which is the motion of carriers in response to an electric field, and diffusion, which is motion in response to a concentration gradient.[3] In either case, the carriers achieve an average directed velocity that is superimposed on their random thermal motion.

With the application of an electric field, the carriers assume an average drift velocity. Under low-field conditions, the drift is a small perturbation on the random thermal motion so the drift velocity is proportional to the electric field strength. Positively charged holes move in the same direction as the electric field but negatively charged electrons move in the opposite direction. The drift current densities are

$$\vec{J}_{n,drift} = q\mu_n n\vec{E} \tag{2.23}$$

and

$$\vec{J}_{p,drift} = q\mu_p p\vec{E}\,, \tag{2.24}$$

where q is the electronic charge, μ_n and μ_p are the electron and hole mobilities, respectively, n and p are the electron and hole concentrations, respectively, and E is the electric field strength. Comparison with Ohm's law shows that the conductivity ($\Omega^{-1}cm^{-1}$) is given by

$$\sigma = q\left(\mu_n n + \mu_p p\right) \tag{2.25}$$

and the resistivity (Ωcm) is given by

$$\rho = \frac{1}{q\left(\mu_n n + \mu_p p\right)}\,. \tag{2.26}$$

The carrier mobilities are a function of temperature and doping concentration. The 300 K mobilities in Si and GaAs are shown in Figure 2.8

Diffusion currents result from the random thermal motion of carriers in the presence of concentration gradients. For either type of carrier, the diffusion

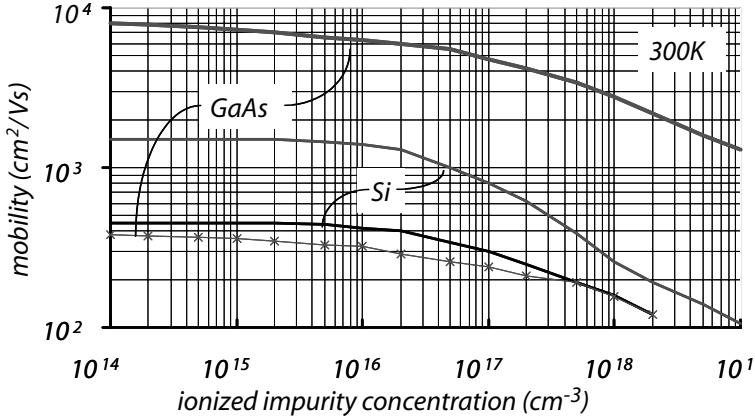


FIGURE 2.8
Electron and hole mobilities in Si and GaAs vs. the ionized impurity (doping) concentration.

particle current is down the gradient. For positively charged holes, the resulting current density is also down the gradient. For negatively charged electrons, the resulting current is up the gradient. The diffusion currents are given by

$$\vec{J}_{n,diff} = qD_n\left(\frac{\partial n}{\partial x}\hat{x} + \frac{\partial n}{\partial y}\hat{y} + \frac{\partial n}{\partial z}\hat{z}\right) \tag{2.27}$$

and

$$\vec{J}_{p,diff} = -qD_p\left(\frac{\partial p}{\partial x}\hat{x} + \frac{\partial p}{\partial y}\hat{y} + \frac{\partial p}{\partial z}\hat{z}\right), \tag{2.28}$$

where D_n and D_p are the diffusivities of electrons and holes, respectively. If only gradients and currents in the x-direction are considered, the one-dimensional equations are

$$J_{n,diff} = qD_n\frac{dn}{dx} \tag{2.29}$$

and

$$J_{p,diff} = -qD_p\frac{dp}{dx}. \tag{2.30}$$

The scattering mechanisms associated with drift and diffusion are similar and involve random thermal motion. Therefore, the mobilities and diffusivities are directly related by the Einstein relationships as follows:

$$D_n = \mu_n\frac{kT}{q} \tag{2.31}$$

and

$$D_p = \mu_p \frac{kT}{q}, \tag{2.32}$$

where k is the Boltzmann constant, T is the absolute temperature, and q is the electronic charge. The quantity kT/q is called the "thermal voltage" and has a value of about 26 mV at a temperature of 300 K.

The vector current densities are found by adding the drift and diffusion components. In one dimension,

$$J_n = q\mu_n nE + qD_n \frac{dn}{dx} \tag{2.33}$$

and

$$J_p = q\mu_p pE - qD_p \frac{dp}{dx}. \tag{2.34}$$

2.7 Carrier Continuity Equations

Continuity equations are used to describe the time rate of change for the minority carrier concentration at a particular point in the semiconductor. For electrons in a p-type semiconductor, the continuity equation is[3]

$$\frac{\partial n_p}{\partial t} = G - \frac{n_p'}{\tau_n} + \frac{1}{q}\nabla \vec{J}_n, \tag{2.35}$$

where n_p' is the excess electron concentration, G is the generation rate for electron–hole pairs, q is the electronic charge, and $\nabla \vec{J}_n$ is the divergence of the electron current density. The one-dimensional form of the continuity equation is

$$\frac{\partial n_p}{\partial t} = G - \frac{n_p'}{\tau_n} + \frac{1}{q}\frac{\partial J_n}{\partial x}, \tag{2.36}$$

where the electron current density may include drift and diffusion components and is given by

$$J_n = q\mu_n nE + qD_n \frac{dn}{dx}. \tag{2.37}$$

The one-dimensional continuity equation is therefore

$$\frac{\partial n_p}{\partial t} = G - \frac{n_p'}{\tau_n} + D_n \frac{\partial^2 n_p}{\partial x^2} + \mu_n \frac{\partial}{\partial x}\left(n_p E\right). \tag{2.38}$$

In devices, electron–hole generation and divergence of the drift current can often be neglected, resulting in the simple form,

$$\frac{\partial n_p}{\partial t} = -\frac{n_p'}{\tau_n} + D_n \frac{\partial^2 n_p}{\partial x^2}. \tag{2.39}$$

For holes in an n-type semiconductor, the continuity equation is

$$\frac{\partial p_n}{\partial t} = G - \frac{p_n'}{\tau_p} + \frac{1}{q}\nabla \vec{J}_p. \tag{2.40}$$

The one-dimensional form is

$$\frac{\partial p_n}{\partial t} = G - \frac{p_n'}{\tau_p} - D_p \frac{\partial^2 p_n}{\partial x^2} + \mu_p \frac{\partial}{\partial x}\left(p_n E\right), \tag{2.41}$$

and the simplified one-dimensional form (neglecting generation and the divergence of the drift current) is

$$\frac{\partial p_n}{\partial t} = -\frac{p_n'}{\tau_p} - D_p \frac{\partial^2 p_n}{\partial x^2}. \tag{2.42}$$

The continuity equations are the starting point for analysis of minority carrier devices such as p–n junctions and bipolar junction transistors.

2.8 Poisson's Equation

Poisson's equation is one of the most important equations governing the operation of semiconductor devices. It is derived from Maxwell's first equation, also known as Gauss's law, which states that the surface integral of the normal component of the electric flux density over any closed surface equals the charge enclosed. In one dimension, Poisson's equation relates the spatial variation of the potential and the electric field intensity with the space charge. The one-dimensional form of Poisson's equation is

$$\frac{d^2\psi_i}{dx^2} = \frac{dE}{dx} = -\frac{\rho}{\varepsilon}, \tag{2.43}$$

where ρ is the space charge density, ε is the permittivity, E is the electric field intensity at the position x, and ψ_i is the electric potential relative to the intrinsic Fermi level. The space charge in a semiconductor is due to mobile charges (electrons and holes) and fixed charges (ionized acceptors and donors). Thus, Poisson's equation may be written as

$$\frac{d^2\psi_i}{dx^2} = \frac{dE}{dx} = -\frac{q}{\varepsilon}\left(N_d - N_a + p - n\right). \tag{2.44}$$

2.9 Dielectric Relaxation Time

In contrast to the minority carrier lifetime, the response time for majority carriers is very short. As a consequence, majority carrier devices (Schottky diodes and FETs) are inherently faster than minority carrier devices (p–n junctions and BJTs). The time constant for the response of majority carriers is the dielectric relaxation time, given by

$$\tau_D = \frac{\varepsilon}{\sigma}, \tag{2.45}$$

where ε is the permittivity of the semiconductor and σ is the conductivity of the semiconductor.

2.10 Summary

Semiconductors are single crystal materials in which the electrical conductivity can be controlled by doping with small quantities of impurities. The periodic arrangement of atoms in the semiconductor gives rise to a unique type of energy band structure in which a valence band and conduction band are separated by a forbidden energy gap.

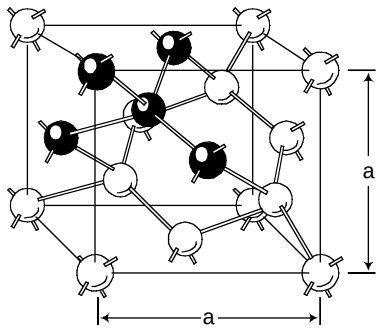
Conduction in a semiconductor is by electrons and holes. The addition of donor impurities increases the electron concentration, resulting in n-type material. The addition of acceptor impurities creates p-type material by increasing the hole concentration. Electrons and holes move by drift and diffusion; drift is in response to an applied electric field and is described by the carrier mobilities, while diffusion is in response to concentration gradients and is described by the diffusivities.

The response time for minority carriers in a semiconductor is called the minority carrier lifetime. The response time for majority carriers in the semiconductor is called the dielectric relaxation time. Majority carriers respond much faster than minority carriers, making majority carrier devices inherently fast.

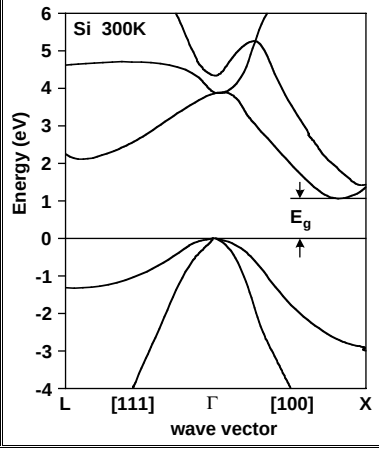
The equations governing semiconductor devices are derived using the current density, continuity, and Poisson equations. Continuity equations are used to describe the time rate of change for the minority carrier concentration at a particular point in the semiconductor. The spatial variation of the potential and electric field intensity in a semiconductor are described by the Poisson equation.

SEMICONDUCTOR MATERIALS QUICK REFERENCE

Diamond Crystal Structure (Silicon)



Band Structure (Silicon)



Nearly all digital integrated circuits are fabricated using single crystal silicon. Silicon crystallizes in the diamond structure with a lattice constant of 5.4310Å at 300K. Pure silicon has a band gap of 1.12 eV at 300K. Its conductivity can be controlled by the addition of dopants to make the material n-type or p-type.

n-type semiconductor

$$n = N_d \qquad p = \frac{n_i^2}{N_d}$$

p-type semiconductor

$$p = N_a \qquad n = \frac{n_i^2}{N_a}$$

Diffusion and Drift Current Density Equations

$$J_n = q\mu_n nE + qD_n \frac{dn}{dx} \qquad J_p = q\mu_p pE - qD_p \frac{dp}{dx}$$

Poisson Equation

$$\frac{d^2\psi_i}{dx^2} = \frac{dE}{dx} = -\frac{q}{\varepsilon}(N_d - N_a + p - n)$$

Continuity Equations

$$\frac{\partial n_p}{\partial t} = -\frac{n_p'}{\tau_n} + D_n \frac{\partial^2 n_p}{\partial x^2} \qquad \frac{\partial p_n}{\partial t} = -\frac{p_n'}{\tau_p} - D_p \frac{\partial^2 p_n}{\partial x^2}$$

Lifetime

$$\tau_p = \frac{p'}{-\frac{dp'}{dt}}$$

f = 10^{-15} p = 10^{-12} n = 10^{-9} μ = 10^{-6} m = 10^{-3} k = 10^{3} M = 10^{6} G = 10^{9}

Problems

P2.1. Determine the room temperature carrier concentrations and resistivity for silicon doped with phosphorus to a concentration of 10^{17} cm^{-3}.

P2.2. Determine the room temperature carrier concentrations and resistivity for silicon doped with boron to a concentration of 10^{17} $cm^{-3.}$

P2.3. Compare the resistivities for intrinsic (undoped) silicon and gallium arsenide at room temperature. Is either of these materials suitable as an insulating substrate for integrated circuits? (Assume that a resistivity greater than 10 MΩcm is required.)

P2.4. CMOS integrated circuits require p-type silicon substrates with a resistivity of 20 Ωcm. Specify the dopant and its required concentration.

References

1. Kittel, C., *Introduction to Solid State Physics*, 7th ed., John Wiley & Sons, New York, 1996.
2. Sze, S.M., *Semiconductor Devices: Physics and Technology*, 2nd ed., John Wiley & Sons, New York, 2001.
3. Streetman, B.G. and Banerjee, S., *Solid State Electronic Devices*, Prentice Hall, Englewood Cliffs, NJ, 1999.
4. Taur, Y. and Ning, T.H., *Fundamentals of Modern VLSI Devices*, Cambridge University Press, New York, 1998.
5. Ghandhi, S.K., *The Theory and Practice of Microelectronics*, Robert E. Krieger Publishing Co., Malabar, FL, 1968.
6. Grove, A.S., *Physics and Technology of Semiconductor Devices*, John Wiley & Sons, New York, 1967.

3

Diodes

3.1 Introduction

P–n junctions are used in some form in all digital integrated circuit families. Occasionally these diodes serve no function other than to isolate the various transistors on a chip. In other cases, the diodes provide protection from electrostatic discharge but do not play an active role in the circuit function. In still other cases, the p–n diodes play an important role in the logic function as well as providing device isolation and input protection.

Besides being important in their own right, p–n junctions are often placed in close proximity within the silicon slice to create transistor action. The resulting bipolar junction transistors are the key components in some logic families including transistor–transistor logic (TTL) and emitter-coupled logic (ECL). This direct relationship makes the p–n junction a good starting point for understanding bipolar transistors, which will be covered in detail in Chapter 4.

P–n junctions are formed by placing an n-type semiconductor in contact with a p-type semiconductor. In practice, this is done by epitaxy, diffusion, or ion implantation. In the first case, a doped epitaxial layer is grown on top of a substrate wafer that has the opposite conductivity type. In the cases of diffusion or ion implantation, a localized region of doped semiconductor is heavily counterdoped to create a junction.

Figure 3.1 shows the simplified construction of a p–n junction and its circuit symbol; this junction acts as a rectifying switch. Its usefulness in digital integrated circuits stems from this rectifying property, which is a result of the change in conductivity type across the junction. The p-type side, or anode, has a large volume concentration of holes but is nearly devoid of electrons.

On the other hand, the n-type side, or cathode, has many electrons but few holes. Under the appropriate bias conditions, holes are readily injected from the anode to the cathode, but not in the opposite direction. Similarly, electrons can be readily injected from the cathode to the anode but not from the anode to cathode. The injection of either type of carrier from the side of high concentration to the side of low concentration results in a conventional

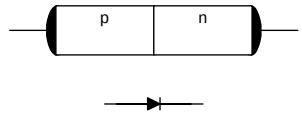


FIGURE 3.1
P–n junction diode and its circuit symbol.

current flowing from anode to cathode. Only very small leakage currents can flow in the opposite direction, except under conditions of reverse breakdown. Therefore, the diode is rectifying and will flow large currents only in the direction of the arrow in the symbol.

This chapter will review the basic properties of p–n junctions, including the zero-bias case, forward and reverse bias, high-frequency small-signal behavior, and large-signal switching behavior. The properties of metal–semiconductor diodes, used extensively in modern bipolar integrated circuits, will also be reviewed. SPICE models for both types of diodes will be presented, as well as physical designs for integrated diodes of both kinds.

3.2 Zero Bias (Thermal Equilibrium)

Under the condition of zero bias, or thermal equilibrium, the p–n junction exhibits separation of charge and a built-in voltage. This is because the placement of the p-type semiconductor in contact with the n-type semiconductor results in large concentration gradients for holes and electrons. The electrons diffuse from the n-type side to the p-type side, where they recombine with the majority carrier holes. Similarly, the holes diffuse down their concentration gradient to the n-type side, where they recombine with the plentiful electrons. This process creates a "depletion region" nearly depleted of free carriers that contains net space charge because of this depletion. The removal of mobile holes from the p-type semiconductor uncovers the immobile, negatively charged ionized acceptors. On the other hand, the depletion of mobile electrons from the n-type side of the junction leaves behind immobile, positively charged donor ions. This situation results in the separation of charge and produces a built-in electric field pointing from the positive charges on the n-type side to the negative charges on the p-type side. Once set up, the built-in field promotes drift currents that oppose the diffusion currents so that a balance is struck in thermal equilibrium.

Consider an abrupt junction in which the doping changes abruptly at the junction but is constant on either side. The band diagram for such an abrupt junction is shown in Figure 3.2. In the neutral n-type region, the Fermi level, E_f, is close to the conduction band edge, E_C. In the neutral p-type region, the Fermi level is close to the valence band edge, E_V. In the depletion region, band bending occurs as described previously. The Fermi level is constant in

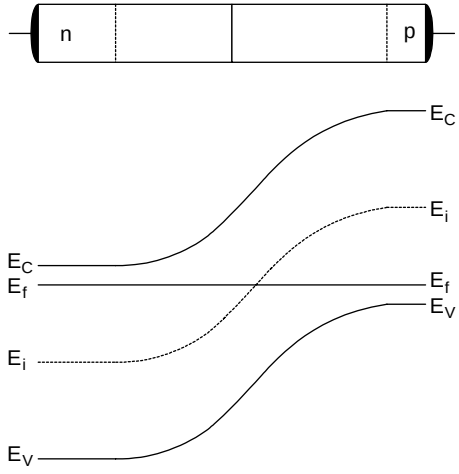


FIGURE 3.2
Equilibrium band diagram for a p–n junction.

equilibrium, which fixes the amount of band bending and also the built-in voltage. The intrinsic Fermi level, E_i, represents the position of the Fermi level for intrinsic (undoped) material and is approximately midway between the band edges.

If the p-type side is doped uniformly with acceptors to a concentration of N_a, then the equilibrium carrier concentrations in the bulk of the p-type semiconductor are

$$\overline{p}_p = N_a \tag{3.1}$$

and

$$\overline{n}_p = \frac{n_i^2}{N_a}. \tag{3.2}$$

Similarly, for the bulk n-type semiconductor doped uniformly with a donor concentration of N_d,

$$\overline{n}_n = N_d \tag{3.3}$$

and

$$\overline{p}_n = \frac{n_i^2}{N_d}. \tag{3.4}$$

The depletion region, also known as the space charge region, may be considered to be completely depleted of free carriers (electrons and holes). Therefore, the space charge density on the p-side of the depletion region is

$$\rho = -qN_a, \tag{3.5}$$

where ρ is the space charge density in C/cm^3, $q = 1.6 \times 10^{-19}$ C, and N_a is the acceptor concentration in cm^{-3}. In similar fashion, the space charge density on the n-side of the depletion region is

$$\rho = qN_d . \tag{3.6}$$

Therefore, the space charge density is a rectangular function of x.

The built-in electric field may be determined using Poisson's equation, which is a modified form of Gauss' law stating that

$$\frac{dE}{dx} = -\frac{\rho}{\varepsilon_s} , \tag{3.7}$$

where E is the electric field intensity, ρ is the space charge density, and ε_s is the permittivity of the semiconductor. The electric field is found by integrating the space charge density and is therefore triangular within the depletion layer and zero outside it.

The electric potential may be found by an additional integration. This is because

$$\frac{dV}{dx} = -E , \tag{3.8}$$

where V is the electric potential. Therefore,

$$V = -\int_0^x E dk , \tag{3.9}$$

which results in a parabolic function of x. The potential difference across the depletion region in thermal equilibrium is called the built-in voltage. This voltage can also be determined from the amount of band bending in the junction. Thus,

$$V_{bi} = \frac{kT}{q} \ln\left(\frac{N_a N_d}{n_i^2}\right), \tag{3.10}$$

where V_{bi} is the built-in voltage, k is the Boltzmann constant, q is the electronic charge, N_a is the acceptor doping on the p-type side, N_d is the donor doping on the n-type side, and n_i is the intrinsic carrier concentration.

The width of the depletion region can be determined by combining Equation 3.9 with Equation 3.10 and using the condition for overall device charge neutrality:

$$x_n N_d = x_p N_a \,. \tag{3.11}$$

The depletion width on the p-type side is

$$x_p = \sqrt{\frac{2\varepsilon_s V_{bi}}{q}\left(\frac{N_d/N_a}{N_a + N_d}\right)} \tag{3.12}$$

and the depletion width on the n-type side is

$$x_n = \sqrt{\frac{2\varepsilon_s V_{bi}}{q}\left(\frac{N_a/N_d}{N_a + N_d}\right)}\,; \tag{3.13}$$

the total depletion width is

$$W = x_p + x_n = \sqrt{\frac{2\varepsilon_s V_{bi}}{q}\left(\frac{1}{N_a} + \frac{1}{N_d}\right)}\,. \tag{3.14}$$

The depletion layer exhibits a capacitance the same as that of a parallel plate capacitor with a separation W:

$$C_T = \frac{\varepsilon_s A}{W} = A\sqrt{\frac{q\varepsilon_s}{2V_{bi}}\left(\frac{1}{N_a} + \frac{1}{N_d}\right)^{-1}}\,. \tag{3.15}$$

Figure 3.3 illustrates the case of the abrupt p–n junction in equilibrium. The space charge density ρ is a rectangular function of x, the electric field E is a triangular function of x, and the electric potential V is a parabolic function of x.

Often p–n junctions are one-sided, that is, one side is doped much more heavily than the other. In an n^+–p junction, the n-type side is doped more heavily. Here, the depletion width exists mostly on the lightly doped side so that

$$W = \sqrt{\frac{2\varepsilon_s V_{bi}}{qN_a}} \quad (n^+\text{–p one-sided junction}) \tag{3.16}$$

and

$$C_T = A\sqrt{\frac{q\varepsilon_s N_a}{2V_{bi}}} \quad (n^+\text{–p one-sided junction}). \tag{3.17}$$

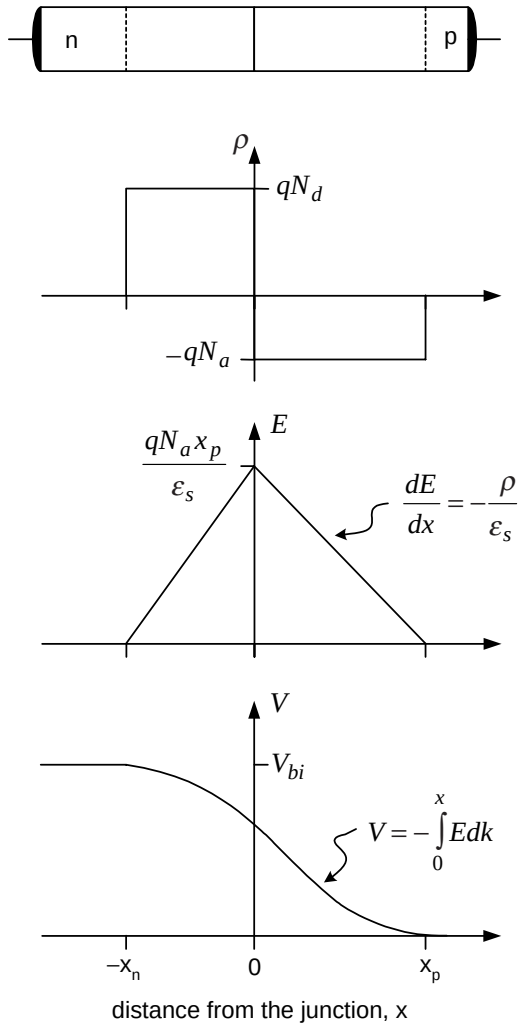


FIGURE 3.3
P–n junction in equilibrium (zero bias).

Example 3.1

Determine the built-in potential, depletion width, and zero-bias depletion capacitance for an n^+–p Si diode at 300 K, with $N_d = 10^{18}$ cm^{-3}, $N_a = 10^{16}$ cm^{-3}, and a junction area of 10^{-5} cm^2.

Solution. The built-in potential is

$$V_{bi} = \frac{kT}{q}\ln\left(\frac{N_a N_d}{n_i^2}\right) = (0.0259\text{ V})\ln\left(\frac{(10^{16}\text{ cm}^{-3})(10^{18}\text{ cm}^{-3})}{(1.45\times10^{10}\text{ cm}^{-3})^2}\right) = 0.816\text{ V}.$$

The depletion width is almost entirely on the p-type side and is

$$W = \sqrt{\frac{2\varepsilon_s V_{bi}}{qN_a}} = \sqrt{\frac{2(11.9)\left(8.85\times10^{-14}\ \text{F/cm}\right)(0.816\ \text{V})}{\left(1.602\times10^{-19}\ \text{C}\right)\left(10^{16}\ \text{cm}^{-3}\right)}}$$

$$= 0.33\times10^{-4}\ \text{cm} = 0.33\ \mu\text{m}.$$

The zero-bias depletion capacitance is

$$C_T = \frac{\varepsilon_s A}{W} = \frac{(11.9)\left(8.85\times10^{-14}\ \text{F/cm}\right)\left(10^{-5}\ \text{cm}^2\right)}{0.33\times10^{-4}\ \text{cm}} = 0.32\times10^{-12}\ \text{F} = 0.32\ \text{pF}\ .$$

3.3 Forward Bias

Under forward-bias conditions (p-type anode at a positive voltage with respect to the n-type cathode), a large current can flow from the anode to the cathode and its magnitude increases exponentially with the applied bias. Here, forward bias will be considered as a small departure from thermal equilibrium and a model will be developed for the current vs. voltage characteristic.[1]

Consider an n^+–p diode in which the conduction is dominated by the injection of electrons into the p-type base. Under thermal equilibrium conditions,

$$\overline{n}_p = \overline{n}_n \exp\left(-\frac{qV_{bi}}{kT}\right), \tag{3.18}$$

where $\overline{n}_p$ is the equilibrium concentration of electrons (minority carriers) on the p-type side and $\overline{n}_n$ is the equilibrium concentration of electrons (majority carriers) on the n-type side. If it is assumed that forward bias is a small departure from thermal equilibrium, then

$$n_p(x=0) = \overline{n}_n \exp\left(-\frac{q(V_{bi}-V)}{kT}\right), \tag{3.19}$$

where $n_p(x = 0)$ is the concentration of minority carriers at the edge of the depletion region on the p-type side and V is the applied bias. Combining these equations yields

$$n_p(x=0) = \overline{n}_p \exp\left(\frac{qV}{kT}\right), \tag{3.20}$$

which is known as the *law of the junction*. Therefore, under forward-bias conditions, minority carrier electrons are injected into the *quasi-neutral* base region; their concentration at the edge of the depletion region is an exponential function of the applied bias.

The current–voltage characteristic for the n^+–p diode can be developed by solution of the continuity equation in the p-type base region. If generation and carrier drift are neglected in the p-type base, then the one-dimensional continuity equation is

$$\frac{\partial n_p}{\partial t} = -\frac{n_p'}{\tau_n} + D_n \frac{\partial^2 n_p}{\partial x^2}, \tag{3.21}$$

where n_p' is the excess minority carrier concentration, given by

$$n_p' = n_p - \overline{n}_p . \tag{3.22}$$

The boundary conditions are

$$n_p'(x=0) = \overline{n}_p\left(e^{qV/kT} - 1\right) \tag{3.23}$$

at the edge of the depletion layer and

$$n_p'(x = W_B) = 0 , \tag{3.24}$$

at the contact, where W_B is the width of the p-type *base* of the diode. The solution is

$$n_p'(x) = \overline{n}_p\left(e^{qV/kT} - 1\right)\frac{\sinh\left(\dfrac{W_B - x}{L_n}\right)}{\sinh\left(\dfrac{W_B}{x}\right)}, \tag{3.25}$$

where $L_n = \sqrt{D_n \tau_n}$ is called the *minority carrier diffusion length*. This behavior is shown in Figure 3.4.

The resulting current that flows is entirely due to the diffusion of minority carriers under moderate-bias conditions. Thus, at the edge of the depletion layer in the p-type base,

$$J_n(x=0) = qD_n \left.\frac{\partial n_p'}{\partial x}\right|_{x=0} = \frac{qD_n \overline{n}_p}{L_n \tanh\left(W_B/L_n\right)}\left(e^{qV/kT} - 1\right). \tag{3.26}$$

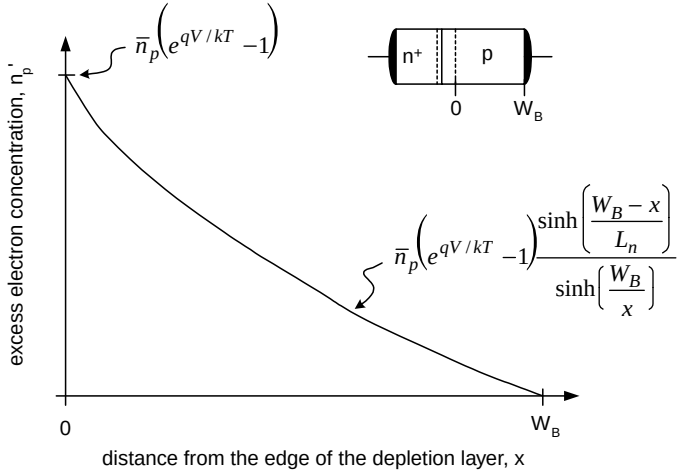


FIGURE 3.4
The excess minority carrier (electron) concentration vs. distance from the edge of the depletion layer for an n⁺–p diode under forward bias. (The injection of holes into the n-type side is negligible for an n⁺–p diode.)

Multiplying by the junction area yields the current

$$I_n = \frac{qAD_n\bar{n}_p}{L_n\tanh\left(W_B/L_n\right)}\left(e^{qV/kT}-1\right) = \frac{qAD_n n_i^2}{L_n N_a\tanh\left(W_B/L_n\right)}\left(e^{qV/kT}-1\right). \quad (3.27)$$

This is the diode equation, which can also be written as

$$I = I_S\left(e^{qV/kT}-1\right), \quad (3.28)$$

where the reverse saturation current I_S for an n⁺–p diode is given by

$$I_S = \frac{qAD_n n_i^2}{L_n N_a\tanh\left(W_B/L_n\right)}. \quad (3.29)$$

The diode equation can be simplified in certain practical cases as shown in the following subsections.

3.3.1 Short-Base n⁺–p Diode

A short-base n⁺–p diode is one in which the width of the p-type base is much less than the minority carrier diffusion length in that region. Thus

$$W_B << L_n \quad (3.30)$$

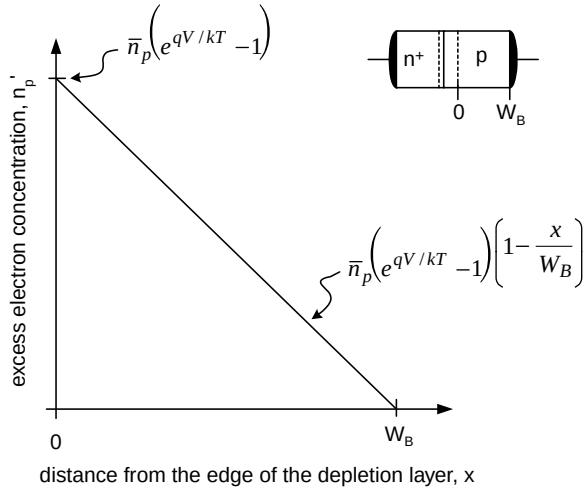


FIGURE 3.5
The excess minority carrier (electron) concentration vs. distance from the edge of the depletion region for a short-base n^+–p diode under forward bias. The profile is linear.

so that the excess minority carrier profile is linear as shown in Figure 3.5. The resulting diffusion current is therefore

$$I_n \approx \frac{qAD_n n_i^2}{N_a W_B}\left(e^{qV/kT}-1\right). \tag{3.31}$$

The short-base diode has high conductance and relatively little stored minority carrier charge under forward bias. Both of these attributes are advantageous in terms of circuit performance.

Example 3.2

Consider an n^+–p diode at 300 K with a junction area of 2×10^{-6} cm^{-2}. For the n^+ emitter, $N_d = 10^{19}$ cm^{-3}, $D_p = 2.0$ cm^{2}s^{-1}, and $\tau = 1$ ns. For the p-type base, $N_a = 10^{16}$ cm^{-3}, $D_n = 15$ cm^{2}s^{-1}, and $\tau_n = 5$ ns. The undepleted base width is 0.2 μm. Determine the forward voltage at a current of 1 mA.

Solution. The forward current is dominated by the injection of electrons into the p-type base because this is an n^+–p diode. The diffusion length for electrons in the base is

$$L_n = \sqrt{D_n \tau_n} = \sqrt{\left(15 \text{ cm}^2\text{s}^{-1}\right)\left(5\times10^{-9} \text{ ns}\right)} = 2.74 \text{ μm}.$$

Because $L_n \gg W_B$, this is a short-base diode. The saturation current is

$$I_S \approx \frac{qAD_n n_i^2}{N_a W_B} = \frac{\left(1.602\times10^{-19}\ \text{C}\right)\left(2\times10^{-6}\ \text{cm}^{-2}\right)\left(15\ \text{cm}^2\text{s}^{-1}\right)\left(1.45\times10^{10}\ \text{cm}^{-3}\right)^2}{\left(10^{16}\ \text{cm}^{-3}\right)\left(0.2\times10^{-4}\ \text{cm}\right)}$$

$$= 5.0\times10^{-15}\ \text{A}.$$

At a current of 1 mA, the forward voltage is therefore

$$V = \frac{kT}{q}\ln\left(\frac{I}{I_S}\right) = (0.026\ \text{V})\ln\left(\frac{10^{-3}\ \text{A}}{5.0\times10^{-15}\ \text{A}}\right) = 0.68\ \text{V}\ .$$

3.3.2 Long-Base n^+–p Diode

A long-base n^+–p diode is one in which the width of the p-type base is much greater than the minority carrier diffusion length in that region. Thus,

$$W >> L_n \tag{3.32}$$

so that

$$I_n \approx \frac{qAD_n n_i^2}{N_a L_n}\left(e^{qV/kT} - 1\right). \tag{3.33}$$

3.4 Reverse Bias

Ideally a p–n junction should block the conduction of current under reverse-bias conditions. In practice, this behavior is only approximated and small leakage currents flow in diodes with moderate reverse bias. With larger reverse-bias voltages applied, it is possible for reverse breakdown, in which the current is limited only by the external circuit, to occur.

3.4.1 Reverse Leakage

Under conditions of moderate bias there are two contributions to the reverse-leakage current. The first is the diffusion current, which may be modeled by the diode equation and is approximately $-I_S$. The second component of the reverse current is the generation current,[2] which is usually dominant in silicon p–n junctions at room temperature. The generation current results from the net generation of electron–hole pairs in the depletion region under

reverse-bias conditions. The carriers are separated by the strong electric field in this region; holes are accelerated toward the p-type anode and electrons are accelerated toward the n-type cathode. The resulting current flows from cathode to anode and is given by

$$I_{gen} = -\frac{qAWn_i}{\tau_{sc}}, \tag{3.34}$$

where q is the electronic charge, A is the junction area, W is width of the depletion region, n_i is the intrinsic carrier concentration, and τ_{SC} is the carrier lifetime in the space charge (depletion) region. Therefore, the generation current is a volume effect. The bias dependence is entirely due to the bias dependence of the depletion width (square root dependence).

The space charge lifetime is related, but not equal, to the minority carrier lifetimes in the quasi-neutral p-type and n-type regions. However, the values of lifetime are shortened by the introduction of *lifetime killer* impurities such as gold. Modern digital bipolar transistors, which are gold-doped for the control of lifetime, also show an associated increase in space charge layer generation and therefore reverse leakage. On the other hand, lifetime controlling impurities are undesirable in CMOS transistors because they do not provide any benefit in terms of speed performance; however, they increase reverse leakage currents and standby power dissipation here as well.

3.4.2 Reverse Breakdown

Under the condition of a sufficiently large reverse bias, breakdown occurs and thus a large reverse current can flow. In fact, the reverse breakdown current is usually limited only by the external circuit. This condition is undesirable in all but a few applications and is prevented by judicious device and circuit design.

Reverse breakdown has two important mechanisms: the avalanche and zener mechanisms. Avalanche breakdown occurs in one-sided n^+–p or p^+–n junctions and involves impact ionization.[3] Zener breakdown involves the quantum mechanical tunneling of electrons through the depletion layer. As such, the zener process can only occur in junctions that have very narrow depletion regions and therefore heavy doping on *both* sides of the junction.

In the case of avalanche breakdown, carriers in the depletion region gain such significant energy from the high electric field that they ionize silicon lattice atoms upon impact. The electron and hole thus created are accelerated by the depletion layer field and will also take part in the process, resulting in an avalanche. For a one-sided n^+–p or p^+–n junction, the avalanche breakdown voltage is only a function of the impurity concentration on the lightly doped side. Qualitatively, a higher impurity concentration will result in a narrower depletion region and a higher peak electric field for a given applied

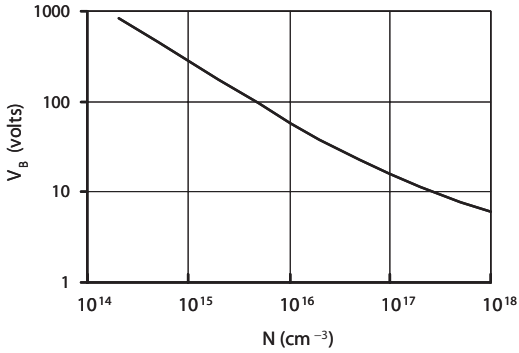


FIGURE 3.6
Reverse breakdown voltage vs. doping concentration for silicon one-sided n^+–p junctions at room temperature.

reverse bias. The ionization coefficients describing the avalanche process are exponential functions of the electric field. Therefore, the breakdown voltage decreases with increasing impurity concentration on the lightly doped side of the junction.

Zener breakdown requires that electrons be able to tunnel through the depletion region under reverse-bias conditions. This necessitates a depletion width on the order of 100 Å (1 Å = 10^{-8} cm) and requires that both sides of the junction be heavily doped.

Figure 3.6 shows the reverse breakdown voltages for silicon one-sided n^+–p junctions at room temperature as a function of N_A. The results are also applicable to one-sided p^+–n junctions as long as the donor concentration is used. In silicon p–n junctions, avalanche breakdown is dominant for breakdown voltages greater than about 8 V, whereas zener breakdown is dominant for breakdown voltages lower than about 4 V. Breakdown in the range between 4 and 8 V takes on a dual character in which the avalanche and zener mechanisms contribute.

3.5 Switching Transients

The large signal transient response of p–n junction diodes is important because of the direct connection to the speed of circuits built using these devices. In this section, the turn-on and turn-off transients for p–n junction diodes will be analyzed; in both cases, the starting point is the charge control equation. It will be shown that the turn-on response is very rapid, but that the turn-off response is considerably slower due to minority carrier charge storage effects.

3.5.1 Charge Control Model

The charge control equation for a one-sided n^+–p diode can be developed by considering the continuity equation for minority carriers in the p-type base. If generation and drift are neglected, this continuity equation is

$$\frac{\partial n_p'}{\partial t} = -\frac{n_p'}{\tau_n} - \frac{1}{qA}\frac{\partial i_n(t)}{\partial x}. \tag{3.35}$$

Multiplying both sides of the equation by $-q$ and integrating over the width of the base region yields

$$-\frac{\partial}{\partial t}\int_0^{W_B} qn_p' dx = \int_0^{W_B} \frac{qn_p'}{\tau_n} dx + \int_0^{W_B} \frac{1}{A}\frac{\partial i_n(t)}{\partial x} dx. \tag{3.36}$$

Multiplying by the junction area yields

$$\frac{dQ_B(t)}{dt} + \frac{Q_B(t)}{\tau_n} - i_n(W_B, t) = -i_n(0, t), \tag{3.37}$$

where Q_B is the excess minority carrier charge stored in the base, in Coulombs. The minority carrier current at the edge of the depletion region includes contributions due to transit and recombination, so

$$-i_n(0,t) = \frac{Q_B}{\tau_n} + \frac{Q_B}{t_{tB}} = \frac{Q_B}{\tau_F}, \tag{3.38}$$

where τ_n is the minority carrier lifetime in the base, t_{tB} is the base transit time for minority carriers, and τ_F is called the "effective forward lifetime." The charge control equation is thus

$$-i_n(0,t) = \frac{dQ_B}{dt} + \frac{Q_B}{\tau_F}. \tag{3.39}$$

If the displacement current in the transition layer capacitance is included, the complete charge control equation is

$$-i_n(0,t) = \frac{dQ_B}{dt} + \frac{Q_B}{\tau_F} - C_T\frac{dv}{dt}. \tag{3.40}$$

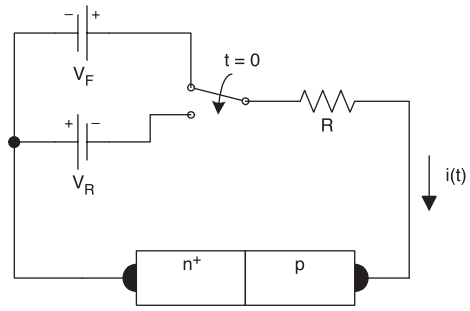


FIGURE 3.7
P–n diode turn-off transient.

3.5.2 Turn-Off Transient

P–n junction diodes require a finite time to turn off due to two effects: storage of excess minority carrier charge and transition layer capacitance. Suppose a reverse voltage is applied after a diode has been forward biased for a long time, as shown in Figure 3.7. Before the bias is switched,

$$i(t) = I_F \approx \frac{V_F}{R}; \quad (t < 0)\,. \tag{3.41}$$

After the bias is switched, a large reverse current flows during the "delay time" and excess minority carriers are removed from the diode:

$$i(t) = -I_R \approx -\frac{V_R}{R}; \quad (0 \le t \le t_s)\,. \tag{3.42}$$

Application of the charge control equation yields

$$I_R \approx \frac{dQ_B}{dt} + \frac{Q_B}{\tau_F} \tag{3.43}$$

and the solution is

$$Q_B(t) \approx I_R\tau_F - \left(I_R + I_F\right)\tau_F \exp\left(-t/\tau_F\right). \tag{3.44}$$

At the end of the storage delay time, the stored minority carrier charge is approximately $-I_R\tau_R$. Here, τ_R is the effective reverse lifetime; it is analogous to τ_F and includes transit and recombination effects. Solving, the storage delay time is

$$t_s = \tau_F \left[\left(1 + \frac{I_F}{I_R} \right) - \left(1 + \frac{\tau_R}{\tau_F} \right) \right]. \tag{3.45}$$

From this analysis it can be concluded that the actual switching speed of the diode depends not only on the device design (through τ_F and τ_R) but also on the circuit design (through I_F and I_R).

3.5.3 Turn-On Transient

Unlike the turn-off transient, the turn-on response is relatively fast and usually does not limit circuit performance. Suppose a p–n junction diode is suddenly forward biased as shown in Figure 3.8. If the diode has been in a state of zero bias for a long time, the diode voltage cannot change abruptly due to the junction capacitance. Thus, the current rises rather abruptly to $I_F \approx V_F/R$ when the switch is thrown at $t = 0$.

After the switch has been closed, the approximate charge control equation may be written as

$$-I_F \approx \frac{dQ_B}{dt} + \frac{Q_B}{\tau_F}. \tag{3.46}$$

The initial and final conditions are

$$Q_B(t = 0) = 0 \tag{3.47}$$

and

$$Q_B(t = \infty) = -I_F \tau_F; \tag{3.48}$$

the solution is

$$Q_B(t) = -I_F \tau_F \left[1 - \exp\left(-t/\tau_F \right) \right]. \tag{3.49}$$

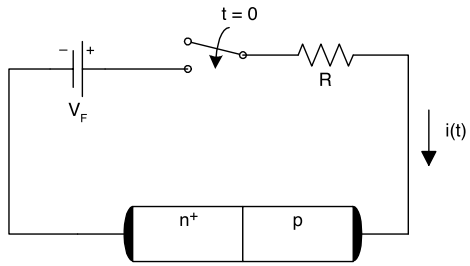


FIGURE 3.8
P–n diode turn-on transient.

The junction voltage as a function of time can be determined if the assumption is that $Q_B(t)$ is proportional to the excess minority carrier concentration at the edge of the depletion region. Then, by the law of the junction,

$$Q_B(t) \propto \overline{n}_p \exp\left(\frac{qv(t)}{kT}\right) \tag{3.50}$$

or

$$v(t) = V_{ss} - \frac{kT}{q} \ln\left(\frac{1}{1+\exp\left(-t/\tau_F\right)}\right), \tag{3.51}$$

where V_{ss} is the steady-state voltage. Therefore, the current and the voltage increase very rapidly when the diode is switched on.

3.6 Metal–Semiconductor Diode

Metal–semiconductor diodes, also known as *Schottky diodes*, are rectifying like p–n junctions.[4] Schottky diodes are formed by depositing metal on an n-type or p-type semiconductor; Figure 3.9 shows the Schottky diode structures schematically with the circuit symbol. In a metal–n-semiconductor diode, the forward-bias current flows from metal to n-semiconductor (the metal acts as the anode). In a metal–p-semiconductor diode, the forward-bias current flows from the p-semiconductor to the metal (the p-semiconductor acts as the anode).

The usefulness of Schottky diodes in digital integrated circuits stems from two unique properties. First, the turn-on voltage is considerably less than for a p–n junction and, second, the Schottky diode is a majority carrier device. The absence of minority carrier storage effects makes Schottky diodes inherently fast. These two characteristics are exploited in Schottky TTL, as will be described in detail in Chapter 5.

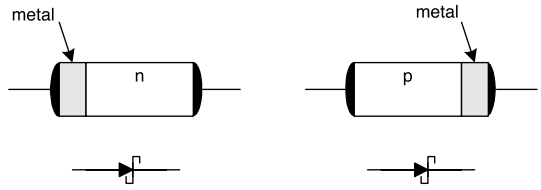


FIGURE 3.9
Metal–semiconductor diodes and circuit symbol.

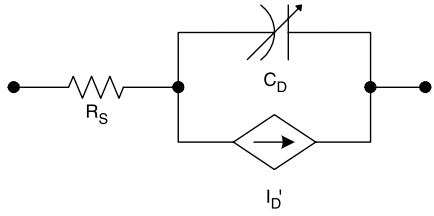


FIGURE 3.10
SPICE circuit model for the p–n junction diode.

3.7 SPICE Models

The SPICE model for the p–n junction diode comprises a Shockley-type current source, a variable capacitance, and a series resistance as shown in Figure 3.10. The current source is modeled by

$$I_D' = IS\left[\exp\left(\frac{qV_D'}{NkT}\right) - 1\right], \tag{3.52}$$

where IS is the reverse saturation current, V_D' is the voltage across the current source, kT/q is the thermal voltage, and N is the *emission coefficient* (sometimes known as the *ideality factor*). If the conduction in the diode is entirely due to diffusion, then the emission coefficient is unity as predicted by the diode equation. At high-current densities, the emission coefficient is greater than one due to drift-aided diffusion under so-called *high-level injection*.

The series resistance is accounted for by

$$V_D = V_D' + I_D RS\,, \tag{3.53}$$

where V_D is the externally applied voltage, I_D is the terminal current, and RS is the series resistance.

The junction capacitance is modeled using the equation

$$C_D = TT\,\frac{IS}{NV_T}\exp\left(\frac{qV_D'}{NkT}\right) + \frac{CJO}{\left(1 - \dfrac{V_D'}{VJ}\right)^M}, \tag{3.54}$$

where the first term is the diffusion capacitance, due to stored excess minority carriers, and the second term is the depletion layer capacitance. Here, TT is the transit time and is equal to τ_F; M is the *grading coefficient* and typically

TABLE 3.1

SPICE Parameters for the p–n Junction Diode Model

Symbol	SPICE Name	Description	Units	Default	Typical
I_S	IS	Saturation current	A	1E-14	2E-15
R_S	RS	Series resistance	Ω	0	2
	N	Emission coefficient	—	1	1
τ_F	TT	Transit time	s	0	1E-10
	CJO	Zero-bias capacitance	F	0	1E-12
V_{bi}	VJ	Built-in voltage	V	1	0.8
m	M	Grading coefficient	—	0.5	0.5

varies between 1/3 and 1/2. For an abrupt junction, $M = 1/2$, as shown previously. Table 3.1 summarizes the parameters used in the SPICE model for the p–n junction diode.

The SPICE model for the Schottky diode is very similar to that for the p–n junction diode — with two important differences. First, the saturation current is orders of magnitude higher than for a p–n junction with the same area; this accounts for the much lower turn-on voltage of the Schottky diode. Second, because of the absence of minority carrier storage, the Schottky diode can be modeled with zero transit time.

3.8 Integrated Circuit Diodes

Diodes that will be placed on integrated circuits have two engineering requirements: they must provide the desired circuit performance and they must be capable of fabrication using the same processing steps as the transistors on the chip. The second requirement is necessary to avoid adding processing steps and increasing the cost. As a consequence, diodes are fabricated differently on MOSFET-based integrated circuits than on BJT-based integrated circuits. Engineers have developed many designs that satisfy these requirements, so no attempt will be made to catalog them here. Instead, some examples of integrated diode structures will be shown.

Integrated p–n junction diodes on bipolar integrated circuits are fabricated using the same processing steps as the npn bipolar transistors; Figure 3.11 shows such a p–n junction diode. In this structure, omission of the emitter results in a one-sided p^+–n^- junction with low series resistance and good breakdown characteristics. (The n^- refers to lightly doped n-type material.)

Integrated Schottky diodes on bipolar integrated circuits are usually of the metal–n-semiconductor type and the metal is usually a silicide such as platinum silicide. Figure 3.12 shows one such Schottky diode compatible with the bipolar process outlined in Chapter 1. This device is surrounded by a p^+ guard ring, which results in good reverse breakdown characteristics. The series resistance is low due to the n^+ buried layer.

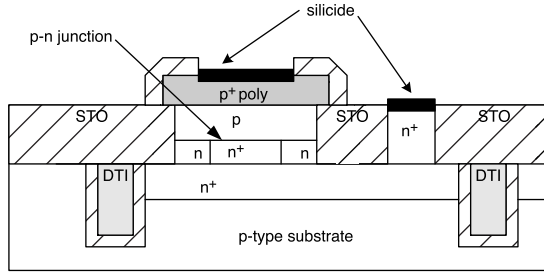


FIGURE 3.11
Integrated p–n junction diode (STO = shallow trench oxide; DTI = deep trench isolation).

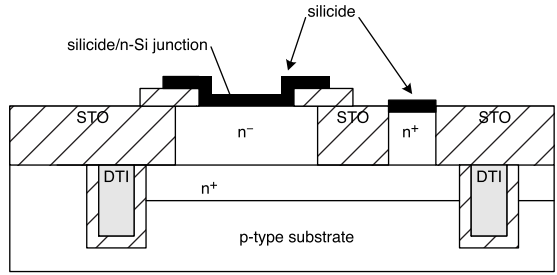


FIGURE 3.12
Integrated Schottky diode (STO = shallow trench oxide; DTI = deep trench isolation).

3.9 PSPICE Simulations

Simulations were performed using PSPICE version 9.1, a Berkeley-type SPICE simulator available for free download.[5]

3.9.1 DC Characteristics

The current vs. voltage (I–V) characteristic was simulated for a diode with a turn-on voltage of 0.7 V using a DC sweep. The turn-on voltage, emission coefficient, and saturation current for a diode are related by

$$V_D = \frac{nkT}{q} \ln\left(\frac{I}{I_S}\right). \tag{3.55}$$

For a diode with a forward voltage of 0.7 V at 1 mA and 300 K, with an emission coefficient of unity, the corresponding saturation current is

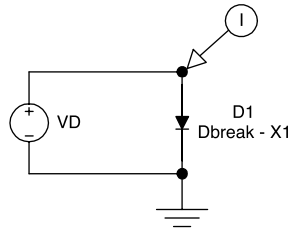


FIGURE 3.13
Circuit used for the simulation of the diode I–V characteristic.

TABLE 3.2

SPICE Parameters Used for Simulation of the Diode I–V Characteristic

Parameter	Value	Units
IS	2.0f	A
N	1.0	—
R	0.1	W

$$I_S = (1\text{mA})\exp\left[\frac{-0.7\text{ V}}{(1)(0.026\text{ V})}\right] = 2.0\times10^{-15}\text{ A}. \tag{3.56}$$

The circuit used appears in Figure 3.13 and diode model parameters used for this simulation are given in Table 3.2.

For the DC sweep, the source voltage, VD, was varied from 0 to 0.8 V with a step of 0.01 V. The resulting I–V characteristic is shown in Figure 3.14. The simulation results show that the approximation $V_D \approx 0.7$ V is justified for values of current on the order of 1 mA.

3.9.2 Effect of Series Resistance

To illustrate the effect of the series resistance, the current vs. voltage characteristics were simulated using the same circuit of Figure 3.13 but with three values of series resistance: 0, 5 Ω, and 10 Ω. The simulation results in Figure 3.15 show that series resistance has little effect on the apparent turn-on voltage for the diode.

3.9.3 Effect of Emission Coefficient

The turn-on voltage of a p–n junction diode is determined by the emission coefficient as well as the saturation current. This is illustrated by the simulated characteristics of Figure 3.16. Device A has N = 1.0 and IS = 2.00 fA;

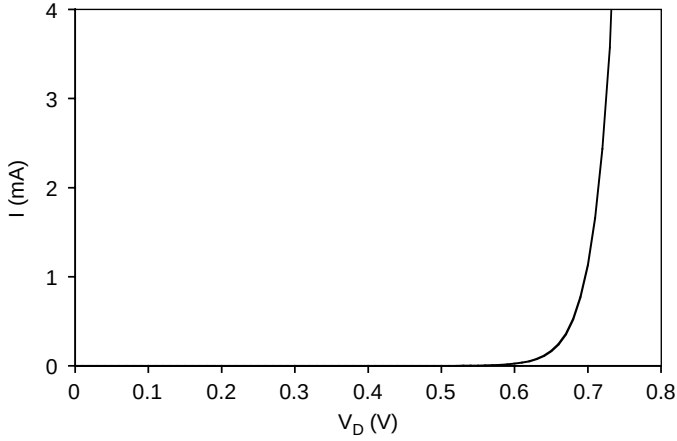


FIGURE 3.14
Simulated diode I–V characteristic.

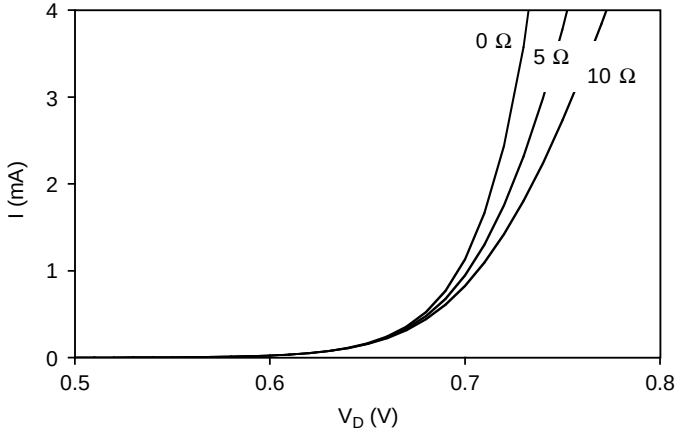


FIGURE 3.15
Current vs. voltage with series resistance as a parameter for a p–n junction diode.

device B has $N = 1.5$ and $IS = 0.160$ pA; and device C has $N = 2.0$ and $IS = 1.42$ nA. All three devices exhibit the same forward voltage at 1 mA but their characteristics have different slopes.

3.9.4 Temperature Behavior

To illustrate the typical temperature behavior for a p–n junction diode, the I–V characteristics were simulated for a 1N4148 diode at several temperatures (Figure 3.17). The current source ID was swept from 1 to 10 mA. A

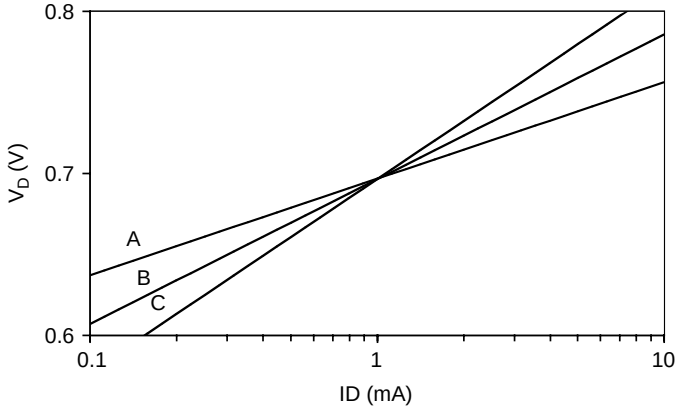


FIGURE 3.16
Voltage vs. current with emission coefficient as a parameter for silicon p–n junction diodes at room temperature. Device A has N = 1.0 and IS = 2.00 fA; device B has N = 1.5 and IS = 0.160 pA; and device C has N = 2.0 and IS = 1.42 nA.

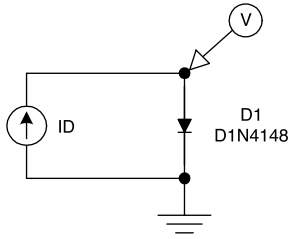


FIGURE 3.17
Circuit for the simulation of the temperature characteristics of a 1N4148 diode.

nested temperature sweep was used, with a start value of 25°C, a finish value of 100°C, and an increment of 25°C.

The simulation results of Figure 3.18 show that, for a given forward current, the diode voltage decreases as the temperature increases. At 25°C and 10 mA, the forward voltage is 0.7258 V; at 75°C and 10 mA, the forward voltage is 0.6451 V. The temperature coefficient for the 1N4148 diode is

$$\frac{\Delta V}{\Delta T} = \frac{0.6451\ \text{V} - 0.7258\ \text{V}}{75^{\circ}\text{C} - 25^{\circ}\text{C}} = -1.6\ \text{mV}/^{\circ}\text{C}\,. \tag{3.57}$$

The negative value of the temperature coefficient indicates that the voltage decreases as the temperature increases. Typically, the forward voltage for a p–n junction has a negative temperature coefficient of approximately –2mV/°C, which is important in emitter-coupled logic gate circuits.

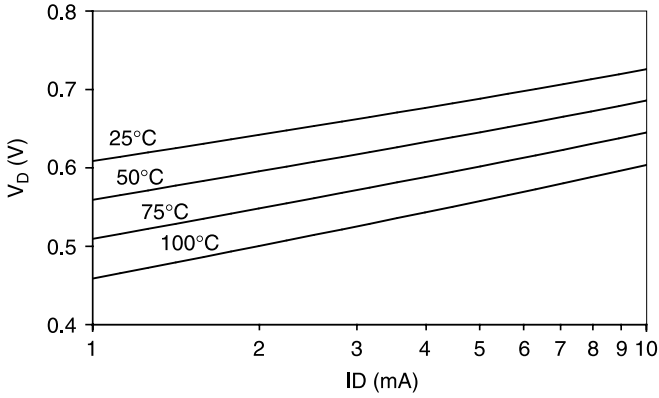


FIGURE 3.18
Simulated voltage vs. current with temperature as a parameter for a 1N4148 diode.

3.10 Summary

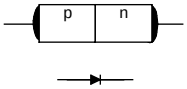
P–n junction diodes are made by placing a p-type semiconductor in contact with an n-type semiconductor. Under equilibrium conditions, a space charge region is near the junction and contains negative space charge on the p-side and positive space charge on the n-side. There is therefore a built-in electric field pointing from the n-side to the p-side and also a built-in voltage. The separation of charge also results in a depletion layer capacitance.

P–n junctions are rectifying, making them useful as switches in digital circuits. With a forward voltage bias from p to n, the current increases exponentially with the applied bias. With a reverse bias from p to n, only a small reverse leakage current flows. The switching behavior of a p–n junction diode may be analyzed using a charge control equation. The results show that the turn-on transient is fast, but the turn-off transient is sluggish due to minority carrier storage effects. Schottky diodes are inherently fast because of the absence of minority carrier storage effects.

P–n junction diodes are modeled in SPICE using a Shockley-type current source, a series resistance, and a parallel capacitance. The parallel capacitance accounts for the sum of the depletion and diffusion capacitances. Integrated circuit diodes are designed to have good circuit performance without unduly adding process complexity. Usually, they are fabricated using the same basic process steps as the transistors fabricated on the same chip.

DIODE QUICK REFERENCE

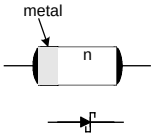
p-n junction diode



$I_S \approx 10^{-15} A$

$V_D \approx 0.7V$

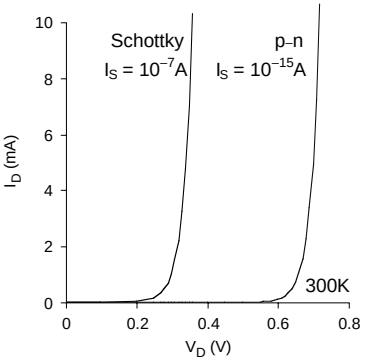
Schottky diode



$I_S \approx 10^{-7} A$

$V_{SBD} \approx 0.3V$

Diode Forward Bias Characteristics



P-n junction diodes are rectifying and can be used as switches in logic gates. Schottky (metal-semiconductor) diodes are also rectifying but are faster switching due to the absence of minority carrier storage effects. Schottky diodes exhibit lower turn-on voltages than p-n junctions (~0.3V versus ~0.7V for silicon devices).

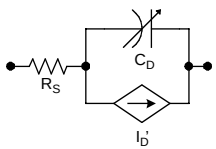
Zero Bias Characteristics

$$V_{bi} = \frac{kT}{q}\ln\left(\frac{N_a N_d}{n_i^2}\right) \qquad W = \sqrt{\frac{2\varepsilon_s V_{bi}}{q}\left(\frac{1}{N_a}+\frac{1}{N_d}\right)} \qquad C_T = \frac{\varepsilon_s A}{W} = A\sqrt{\frac{q\varepsilon_s}{2V_{bi}}\left(\frac{1}{N_a}+\frac{1}{N_d}\right)^{-1}}$$

Forward Bias Characteristics

$$I = I_0\left(e^{qV/kT}-1\right) \qquad I_0 \approx \frac{qAD_n n_i^2}{N_a W_B} \text{ (short-base diode)}$$

SPICE Model



$$I_D' = IS\left[\exp\left(\frac{V_D'}{NV_T}\right)-1\right] \qquad V_D = V_D' + I_D RS$$

$$C_D = TT\,\frac{IS}{NV_T}\exp\left(\frac{V_D'}{NV_T}\right)+\frac{CJO}{\left(1-\frac{V_D'}{VJ}\right)^M}$$

Charge Control Model and Large Signal Switching Behavior

$$-i_n(0,t) = \frac{dQ_B}{dt}+\frac{Q_B}{\tau_F}-C_t\frac{dv}{dt} \qquad t_s = \tau_F\left[\left(1+\frac{I_F}{I_R}\right)-\left(1+\frac{\tau_R}{\tau_F}\right)\right]$$

f = 10^{-15} p = 10^{-12} n = 10^{-9} μ = 10^{-6} m = 10^{-3} k = 10^{3} M = 10^{6} G = 10^{9}

Laboratory Exercises

L3.1. For a commercially available diode, measure the current vs. voltage characteristic. From the measured characteristic, determine the reverse saturation current, emission coefficient, and series resistance. Obtain the SPICE model for the diode from the Internet and compare your model parameters to those provided by the manufacturer.

L3.2. For a commercially available diode, measure the capacitance vs. the reverse bias. From the measured characteristic, determine VJ, CJO, and M. Obtain the SPICE model for the diode from the Internet and compare your model parameters to those provided by the manufacturer.

Problems

P3.1. Consider a p–n junction diode with negligible series resistance. The voltage drop is 0.7 V with a forward current of 5 mA. Determine the reverse saturation current for the case of

1. A unity emission coefficient
2. An emission coefficient of 1.3

P3.2. Consider diodes with unity emission coefficients and negligible series resistance. Determine the reverse saturation current for

1. A p–n junction that exhibits a voltage drop of 0.7 V at a forward current of 1 mA
2. A Schottky diode that exhibits a voltage drop of 0.3 V at a forward current of 1 mA

P3.3. Suppose a diode has a unity emission coefficient. Determine the change in the forward-bias voltage that will cause a 10-fold increase in the forward current.

P3.4. Consider a Si p^+–n junction at 300 K with $N_a = 10^{18}$ cm^{-3} and $N_d = 10^{16}$ cm^{-3}. The junction area is 10^{-5} cm^2. Determine the

1. Built-in potential
2. Zero-bias depletion width
3. Zero-bias depletion capacitance

P3.5. Consider a Si p^+–n junction at 300 K with 2×10^{18} cm^{-3} and $N_d = 10^{16}$ cm^{-3}. The junction area is 3.2×10^{-6} cm^2. Determine and plot $1/C^2$ vs. the reverse-bias voltage.

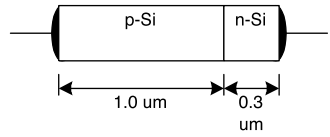


FIGURE 3.19
P–n junction (P3.6).

TABLE 3.3

Material Parameters of p–n Junction Illustrated in Figure 3.19

p-Si	n-Si
$N_a = 10^{15}$ cm^{-3}	$N_d = 10^{18}$ cm^{-3}
$D_n = 25$ cm^{2}s^{-1}	$D_p = 8$ cm^{2}s^{-1}
$\tau_n = 20$ ns	$\tau_p = 3$ ns

P3.6. Consider the p–n junction shown in Figure 3.19. T = 300 K. The material parameters are given in Table 3.3 and the junction area is 10^{-5} cm^2.

1. Determine the reverse saturation current.
2. Determine the bias voltage at a forward current of 1 mA, assuming the emission coefficient is unity.

References

1. Shockley, W., The theory of p–n junctions in semiconductors and p–n junction transistors, *Bell Sys. Tech. J.*, 28, 435, 1949.
2. Sah, C.T., Noyce, R.N., and Shockley, W., Carrier generation and recombination in p–n junctions and p–n junction characteristics, *Proc. IRE*, 45, 1228, 1957.
3. Ghandhi, S.K., *Semiconductor Power Devices*, John Wiley & Sons, New York, 1977.
4. Sze, S.M., *Physics of Semiconductor Devices*, John Wiley & Sons, New York, 1981.
5. www.cadence.com.

4

Bipolar Junction Transistors

4.1 Introduction

Bipolar junction transistors (BJTs) comprise two p–n junctions. If the two junctions are in close proximity, they will interact. Thus, the excess minority carriers injected by a forward-biased junction can be collected by the other junction if it is reverse biased. This results in transistor action, or current gain. BJTs are of two types: npn and pnp. Npn transistors are used more often because of their superior high-frequency and high-power performance. However, pnp transistors are occasionally used when mandated by circuit considerations. Although the discussion here will center on the npn transistor, the pnp transistor is essentially similar except that all voltages and currents have the opposite polarity.

Figure 4.1 shows an npn transistor and its circuit symbol. The p-type region is called the base (abbreviated "B"). The n-type regions are called the emitter and collector (abbreviated "E" and "C," respectively). Except in rare cases, bipolar transistors are not symmetric. Thus the emitter is doped much more heavily than the collector and the emitter and collector are not interchangeable. BJTs may be used as current or voltage amplifiers, for example, in radios, televisions, and wireless phones. In digital circuits, however, BJTs are most often used as switches. In such applications, a signal routed to the base controls the flow of current from the collector to the emitter. References 1 through 10 provide good general information on bipolar junction transistors.

4.2 The Bipolar Junction Transistor in Equilibrium

The equilibrium band diagram for an npn bipolar transistor appears in Figure 4.2. The emitter is degenerately doped so that the Fermi level is coincident with the conduction band edge, E_C, in the neutral emitter. The base is heavily (but not degenerately) doped p-type. Therefore, the Fermi level is close to the valence band edge, E_V, in the neutral base. The depletion

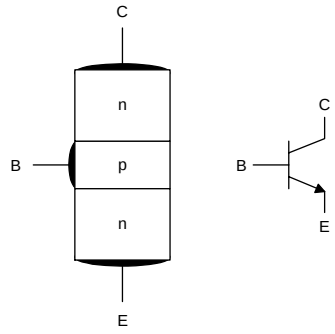


FIGURE 4.1
Npn bipolar transistor and circuit symbol.

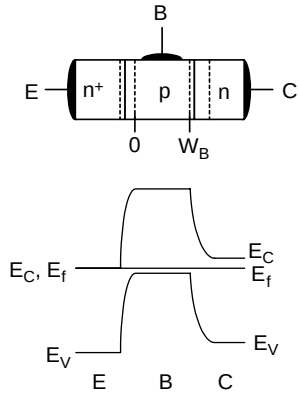


FIGURE 4.2
Equilibrium band diagram for an npn bipolar transistor.

region of the emitter–base junction resides mostly on the more lightly doped side (the base). The collector is moderately doped n-type, so the Fermi level is close to the conduction band edge in the neutral collector. The depletion region of the collector–base junction resides mostly on the collector side because the collector is more lightly doped.

4.3 DC Operation of the Bipolar Junction Transistor

The BJT has four modes of operation:

1. *Cutoff* — both junctions are reverse biased.
2. *Forward active* — the base–emitter junction is forward biased but the base–collector junction is reverse biased.

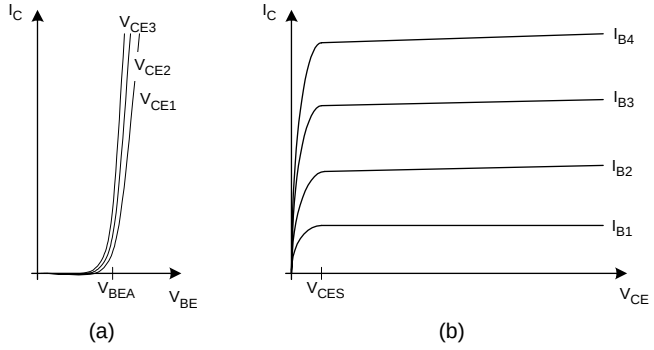


FIGURE 4.3
Npn BJT characteristics: (a) I_C vs. V_{BE} with V_{CE} as a parameter and (b) I_C vs. V_{CE} with I_B as a parameter.

3. *Reverse active* — the base–emitter junction is reverse biased but the base–collector junction is forward biased.
4. *Saturation* — both junctions are forward biased.

Figure 4.3 shows typical characteristics for an npn bipolar junction transistor. Figure 4.3a shows the collector current vs. the base–emitter voltage with the collector–emitter voltage as a parameter. These characteristics show that the base–emitter junction turns on at a voltage of V_{BEA} (typically 0.7 V). In Figure 4.3b, the collector current is plotted as a function of the collector–emitter voltage, with the base current as a parameter. This results in a family of curves, one curve for each value of base current. Forward active operation corresponds to the approximately flat portions of the curves. Saturation corresponds to the sloping parts of the curves for which $V_{CE} \leq V_{CES}$. The cutoff condition occurs with $I_C = 0$ and coincides with the V_{CE} axis. Reverse active operation (not shown) would occur in the third quadrant with $I_C < 0$ and $V_{CE} < 0$.

4.3.1 Cutoff Operation

In the cutoff mode of operation, both junctions are reverse biased (V_{BE} and V_{BC} are negative). Therefore, negligibly small leakage currents flow. For hand calculations, it is assumed that

$$I_C \approx I_E \approx I_B \approx 0 \tag{4.1}$$

for cutoff operation.

4.3.2 Forward Active Operation

In the forward active mode of operation, the base–emitter junction is forward biased but the base–collector junction is reverse biased ($V_{BE} > 0$ but $V_{BC} < 0$).

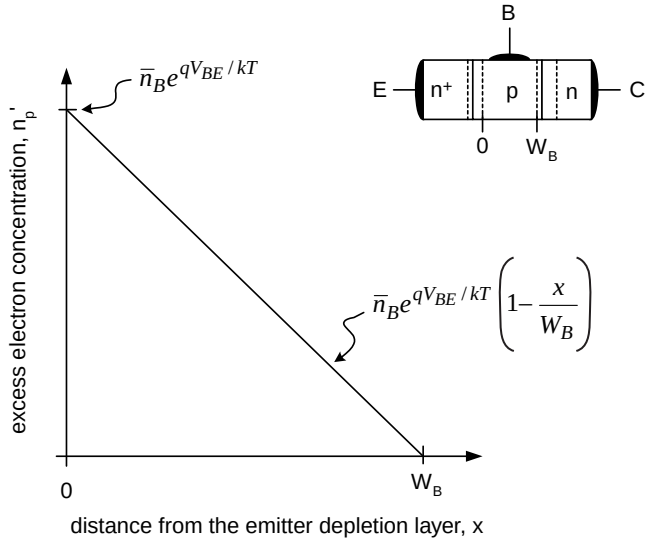


FIGURE 4.4
Excess minority carrier concentration in the base of an npn transistor during forward active operation.

In this mode, the transistor acts like a current-controlled current source, in which the current in the base–emitter diode controls the collector current. The ratio of the collector current to the base current is the *common emitter current gain*, β_F. The emitter injects minority carriers (electrons) into the base. The base is so short that most (~98%) of these injected carriers are able to diffuse to the reverse-biased base–collector junction, where they are collected. During forward active operation, minority carrier charge is stored in the base and the time constant for this charge is τ_F, the forward transit time.

4.3.2.1 Collector Current

For forward active operation of a good transistor, a linear profile for the excess minority carrier concentration is in the base, as shown in Figure 4.4. This is analogous to the situation in a short-base diode and such a transistor is called a short-base transistor.[11]

With a forward bias, V_{BE}, applied between the base and emitter, the excess minority concentration at the emitter end of the base may be determined using the law of the junction, which is

$$n_B(0) - \overline{n}_B \approx \overline{n}_B\left(e^{qV_{BE}/kT} - 1\right), \tag{4.2}$$

where

n_B = electron concentration in the base and

$\overline{n}_B$ = equilibrium electron concentration in the base.

The excess minority carrier concentration at the collector end of the base is zero:

$$n_B(W_B) - \bar{n}_B \approx 0 \,. \tag{4.3}$$

Solving the continuity equation subject to these boundary conditions yields the excess electron concentration in the base as a function of distance:

$$n_B - \bar{n}_B = \bar{n}_B\left(e^{qV_{BE}/kT} - 1\right)\frac{\sinh\left\{(W_B - x)/L_{nB}\right\}}{\sinh(W_B/L_{nB})}\,, \tag{4.4}$$

where L_{nB} = diffusion length for electrons in the base and W_B = base width.

For a short-base transistor $W_B \ll L_{nB}$, so the excess electron concentration is a linear function of distance:

$$n_B - \bar{n}_B \approx \bar{n}_B\left(e^{qV_{BE}/kT} - 1\right)\left(\frac{W_B - x}{W_B}\right). \tag{4.5}$$

Use of the linear function amounts to neglecting the recombination of electrons with holes in the base. Based on this assumption and the fact that $\bar{n}_B = n_i^2/N_A$, the collector current due to the diffusion of minority carriers across the base is

$$I_C = qAD_{nB}\frac{dn_B}{dx} \approx \frac{qAD_{nB}n_i^2}{W_B N_{aB}}\left(e^{qV_{BE}/kT} - 1\right). \tag{4.6}$$

where

q = electronic charge
A = emitter junction area
D_{nB} = diffusivity of electrons in the base
N_i = intrinsic carrier concentration in Si
W_B = base width
N_{aB} = acceptor concentration in the base

Therefore, the collector current increases exponentially with the base–emitter bias voltage.

4.3.2.2 Current Gain

The common base current gain, α_F, of the bipolar transistor is defined as

$$\alpha_F = \frac{\partial I_C}{\partial I_E}\,. \tag{4.7}$$

This can be factored into three terms as follows:

$$\alpha_F = \frac{\partial I_{nE}}{\partial I_E}\frac{\partial I_{nC}}{\partial I_{nE}}\frac{\partial I_C}{\partial I_{nC}} = \gamma_E \alpha_T M\,, \tag{4.8}$$

where

I_E = emitter current
I_{nE} = emitter current due to electron injection into the base
I_C = collector current
I_{nC} = collector current due to collection of electrons from the base
γ_E = emitter injection efficiency
M = collector multiplication factor

The emitter injection efficiency quantifies the fraction of the emitter current that is due to the injection of electrons into the base and is given by

$$\gamma_E = \frac{\partial I_{nE}}{\partial\left(I_{nE} + I_{pE}\right)}, \tag{4.9}$$

where I_{pE} is the component of the emitter current due to the injection of holes into the emitter. However,

$$I_{nE} \approx \frac{qAD_{nB}n_i^2}{W_B N_{aB}}\left(e^{qV_{BE}/kT} - 1\right) \tag{4.10}$$

and

$$I_{pE} \approx \frac{qAD_{pE}n_i^2}{L_{pE}N_{dE}}\left(e^{qV_{BE}/kT} - 1\right), \tag{4.11}$$

where D_{pE} = diffusivity of holes in the emitter, L_{pE} = diffusion length for holes in the emitter, and N_{DE} = donor concentration in the emitter. Therefore, the emitter injection efficiency is given by

$$\gamma_E = \left(1 + \frac{D_{pE}N_{aB}W_B}{D_{nB}N_{dE}L_{pE}}\right)^{-1}. \tag{4.12}$$

The base transport factor quantifies the fraction of the injected electrons that traverse the base without recombining with majority carriers (holes). It is therefore given by

$$\alpha_T = \frac{\left(\partial n_B/\partial x\right)_{x=W_B}}{\left(\partial n_B/\partial x\right)_{x=0}}. \tag{4.13}$$

Accounting for the recombination of electrons in the base, the electron concentration vs. distance in the base is

$$n_B - \overline{n}_B = \overline{n}_B\left(e^{qV_{BE}/kT} - 1\right)\frac{\sinh\left\{\left(W_B - x\right)/L_{nB}\right\}}{\sinh\left(W_B/L_{nB}\right)}; \tag{4.14}$$

therefore,

$$\alpha_T = \left\{\cosh\left(W_B/L_{nB}\right)\right\}^{-1} \approx \left(1 + \frac{W_B^2}{2L_{nB}^2}\right)^{-1}. \tag{4.15}$$

The injected electrons that diffuse across the width of the base are collected at the edge of the collector depletion region by the built-in electric field. This field is directed from the collector toward the base and thus accelerates negatively charged electrons toward the collector region. For sufficiently high values of the collector-to-base bias voltage, some of the collected electrons will gain enough energy to cause impact ionization, resulting in additional current flow. This behavior can be described by the empirical relationship

$$M = \left[1 - \left(\frac{V_{CB}}{BV}\right)^m\right]^{-1}, \tag{4.16}$$

where M = collector multiplication factor, BV = collector–base breakdown voltage, and m = 4 for electrons and 2 for holes.

For normal operation of a digital npn bipolar transistor, the emitter injection efficiency and the collector multiplication factor are close to unity so that

$$\alpha_F \approx \alpha_T \approx \left(1 + \frac{W_B^2}{2L_{nB}^2}\right)^{-1}.$$

The common emitter current gain β_F is defined as

$$\beta_F = \frac{\partial I_C}{\partial I_B} = \frac{\partial I_C}{\partial\left(I_E - I_C\right)} = \frac{\partial I_C/\partial I_E}{1 - \partial I_C/\partial I_E};$$

therefore, the common emitter current gain is related to the common base current gain by

$$\beta_F = \frac{\alpha_F}{1 - \alpha_F}. \tag{4.17}$$

Example 4.1

Estimate β_F for an npn bipolar transistor with a base width of 100 nm. The base doping is 10^{18} cm^{-3} and the base minority carrier lifetime is 0.7 ns. The emitter doping is 10^{20} cm^{-3}, the emitter minority carrier lifetime is 0.2 ns, and the mobility of holes in the emitter is approximately 10 cm^2 V^{-1}s^{-1}.

Solution. If the base doping is 10^{18} cm^{-3}, the electron mobility in the base can be determined from Figure 2.8 to be $\mu_n \approx 250$ cm^2 V^{-1}s^{-1}. Using the Einstein relationship, the diffusivity of electrons in the base is

$$D_{nB} = \frac{kT}{q}\mu_n = (0.026\text{ V})(250\text{ cm}^2\text{V}^{-1}\text{s}^{-1}) = 6.5\text{ cm}^2\text{s}^{-1}.$$

The diffusion length for electrons in the base is

$$L_{nB} = \sqrt{D_{nB}\tau_{nB}} = \sqrt{(6.5\text{ cm}^2\text{s}^{-1})(0.7\times10^{-9}\text{s})} = 6.7\times10^{-5}\text{cm}$$

and the base transport factor is

$$\alpha_T \approx \left(1+\frac{W_B^2}{2L_{nB}^2}\right)^{-1} = \left(1+\frac{(100\text{ nm})^2}{2(670\text{ nm})^2}\right)^{-1} = 0.989.$$

Using the Einstein relationship, the diffusivity of holes in the emitter is

$$D_{pE} = \frac{kT}{q}\mu_{pE} = (0.026\text{ V})(10\text{ cm}^2\text{V}^{-1}\text{s}^{-1}) = 0.26\text{ cm}^2\text{s}^{-1},$$

the diffusion length for holes in the emitter is

$$L_{pE} = \sqrt{D_{pE}\tau_{pE}} = \sqrt{(0.26\text{ cm}^2\text{s}^{-1})(0.2\times10^{-9}\text{s})} = 7.2\times10^{-6}\text{ cm} = 72\text{ nm},$$

and the emitter injection efficiency is

$$\gamma_E = \left(1+\frac{D_{pE}N_{AB}W_B}{D_{nB}N_{DE}L_{PE}}\right)^{-1}$$

$$= \left(1+\frac{(0.26\text{ cm}^2\text{s}^{-1})(10^{18}\text{ cm}^{-3})(100\times10^{-7}\text{ cm})}{(6.5\text{ cm}^2\text{s}^{-1})(10^{20}\text{ cm}^{-3})(72\times10^{-7}\text{ cm})}\right)^{-1} = 0.9994$$

The common base current gain is therefore $\alpha_F \approx \gamma_E \alpha_T = (0.9994)(0.989) = 0.988$ and the common emitter current gain is

$$\beta_F = \frac{\alpha_F}{1-\alpha_F} = \frac{0.988}{1-0.988} = 82 .$$

In this case, the emitter injection efficiency is close to unity, so the current gain is determined primarily by the base transport factor.

4.3.2.3 Transit Time

Diffusing minority carriers in the base of a bipolar transistor take a finite time to cross the base — a *base transit time* that is a fundamental limitation on the speed of the transistor. The base transit time may be determined for a uniform short-base transistor as follows. The excess minority carrier concentration is given by

$$n_B - \overline{n}_B \approx \overline{n}_B\left(e^{qV_{BE}/kT} - 1\right)\left(\frac{W_B - x}{W_B}\right) \tag{4.18}$$

and the resulting diffusion current is given by

$$I_C = qAD_{nB}\overline{n}_B\left(e^{qV_{BE}/kT} - 1\right)\left(\frac{W_B - x}{W_B}\right). \tag{4.19}$$

The effective velocity for a diffusing electron in the base is

$$v = \frac{I_C}{qA\left(n_B - \overline{n}_B\right)} = \frac{D_{nB}}{W_B - x}, \tag{4.20}$$

so the base transit time is

$$t_{tB} = \int_0^{W_B} \frac{dx}{v} = \frac{W_B^2}{2D_{nB}} . \tag{4.21}$$

Integrated bipolar transistors have graded doping in their bases, which can shorten the transit time considerably (to one half the value predicted for the uniform base transistor). Also, this makes the reverse transit time much longer than the forward transit time in digital bipolar transistors.

Example 4.2

Estimate the base transit time for an npn bipolar transistor if the undepleted base width is 100 nm, the base doping is 10^{18} cm^{-3}, and the base minority carrier lifetime is 0.7 ns.

Solution. If the base doping is 10^{18} cm^{-3}, the electron mobility in the base can be determined from Figure 2.8 to be $\mu_n \approx 250$ cm^{2}V^{-1}s^{-1}. Using the Einstein relationship, the diffusivity of electrons in the base is

$$D_{nB} = \frac{kT}{q}\mu_n = (0.026\text{ V})(250\text{ cm}^2\text{V}^{-1}\text{s}^{-1}) = 6.5\text{ cm}^2\text{s}^{-1}$$

and the base transit time is

$$t_{tB} = \frac{W_B^2}{2D_{nB}} = \frac{(100\times10^{-7}\text{ cm})^2}{2(6.5\text{ cm}^2\text{V}^{-1}\text{s}^{-1})} = 7.7\text{ ps}.$$

Therefore, the base transit time is smaller than the propagation delays in emitter-coupled logic circuits.

4.3.2.4 Approximate Analysis

For hand calculations, forward active operation can be modeled using

$$V_{BE} \approx V_{BEA} \tag{4.22}$$

and

$$I_C \approx \beta_F I_B . \tag{4.23}$$

Here β_F is the forward common-emitter current gain and is related to the forward common base current gain by

$$\beta_F = \frac{\alpha_F}{1-\alpha_F} . \tag{4.24}$$

For digital bipolar transistors, typical values are $V_{BEA} = 0.7$ V, $\alpha_F = 0.98$, and $\beta_F = 50$.

4.3.3 Reverse Active Operation

In the reverse active mode of operation, the base–emitter junction is reverse biased but the base–collector junction is forward biased ($V_{BE} < 0$ but $V_{BC} > 0$). In this mode, the transistor also acts like a current-controlled current source,

but for reverse active operation the current in the base–collector diode controls the emitter current. The collector injects minority carriers (electrons), which are collected by the emitter, into the base. Thus the collector and emitter have exchanged roles compared to the forward active mode. The fraction of minority carriers that crosses the base without recombination is α_R, the reverse common-base current gain. The time constant for stored minority carrier charge in the base during reverse active operation is τ_R, the reverse transit time.

For hand calculations, reverse active operation can be modeled using

$$V_{BC} \approx V_{BCA} \tag{4.25}$$

and

$$I_E \approx \beta_R I_B . \tag{4.26}$$

Here, β_R is the reverse common-emitter current gain and is related to the reverse common base current gain by

$$\beta_R = \frac{\alpha_R}{1-\alpha_R} . \tag{4.27}$$

For digital bipolar transistors, typical values are $V_{BCA} = 0.7$ V, $\alpha_R = 0.33$, and $\beta_R = 0.5$.

4.3.4 Saturation Operation

In the saturation mode of operation, both junctions are forward biased ($V_{BE} > 0$ and $V_{BC} > 0$) and inject minority carriers into the base. In addition, significant minority carrier charge is injected from the base into the collector. As a result, a saturated bipolar transistor has a large amount of stored minority carrier charge in the base and collector. For this reason, bipolar transistors are very slow to switch to the cutoff mode once they are allowed to saturate.

The saturated transistor acts like a closed switch, with a small voltage drop from collector to emitter. For hand calculations, saturation operation is modeled using

$$V_{BE} \approx V_{BES} \tag{4.28}$$

and

$$V_{CE} \approx V_{CES} . \tag{4.29}$$

For digital bipolar transistors, typical values are $V_{BES} = 0.8$ V and $V_{CES} = 0.1$ V.

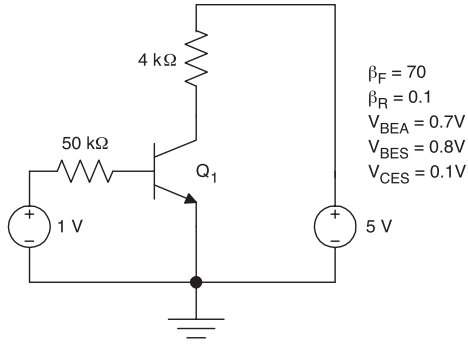


FIGURE 4.5
Bipolar transistor with bias for determining the mode of operation.

Example 4.3

Determine the mode of operation for the transistor of Figure 4.5 and find its collector–emitter voltage.

Solution. Assuming forward active operation,

$$I_B = \frac{1\text{ V} - 0.7\text{ V}}{50\text{ k}\Omega} = 0.006\text{ mA}$$

and $I_C = \beta_F I_B = (70)(0.006\text{ mA}) = 0.42\text{ mA}$. The collector–emitter voltage will be $V_{CE} = V_{CC} - I_C R_C = 5\text{ V} - (0.42\text{ mA})(4\text{ k}\Omega) = 3.3\text{ V}$. This is consistent with forward active operation, so the starting assumption was correct.

Example 4.4

Determine the mode of operation for the transistor of Figure 4.6 and find the collector current.

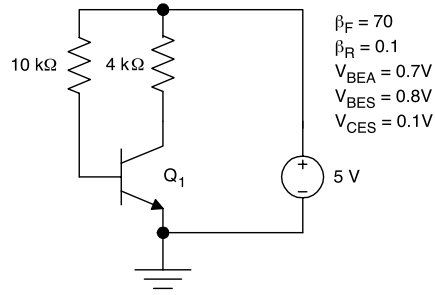


FIGURE 4.6
Example of bipolar transistor with bias.

Solution. Assuming forward active operation,

$$I_B = \frac{5\text{ V} - 0.7\text{ V}}{10\text{ k}\Omega} = 0.43\text{ mA}$$

and $I_C = \beta_F I_B = (70)(0.43\text{ mA}) = 30.1\text{ mA}$. However, the maximum collector current in this circuit is

$$I_C(\max) = \frac{V_{CC} - V_{CES}}{R_C} = \frac{5V - 0.1\text{ V}}{4\text{ k}\Omega} = 1.22\text{ mA}\,.$$

Therefore, the transistor will be saturated with a collector current equal to 1.22 mA.

4.4 Ebers–Moll Model

The DC operation of the bipolar transistor can be modeled using a coupled diode, or Ebers–Moll model,[12] as shown in Figure 4.7. Here the dependent current sources model the transistor action. The equation form of the Ebers–Moll model may be realized by using the diode equation for each of the coupled diodes. This yields the set of equations

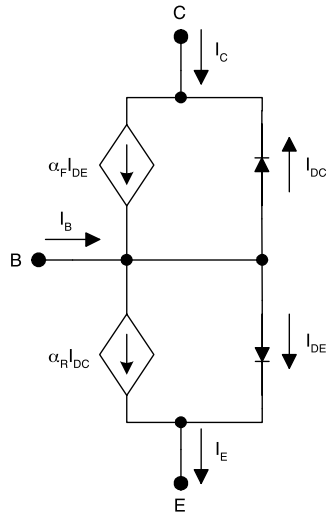


FIGURE 4.7
Ebers–Moll model for DC operation of the bipolar transistor.

$$I_C = I_S\left[\exp\left(\frac{qV_{BE}}{kT}\right)-1\right]-\frac{I_S}{\alpha_R}\left[\exp\left(\frac{qV_{BC}}{kT}\right)-1\right], \tag{4.30}$$

$$I_E = \frac{I_S}{\alpha_F}\left[\exp\left(\frac{qV_{BE}}{kT}\right)-1\right]-I_S\left[\exp\left(\frac{qV_{BC}}{kT}\right)-1\right], \tag{4.31}$$

and

$$I_B = I_E - I_C\,. \tag{4.32}$$

Example 4.5

Estimate the reverse saturation current for an npn bipolar transistor if the undepleted base width is 100 nm, the base doping is 10^{18} cm^{-3}, the base minority carrier lifetime is 0.7 ns, and the emitter area is 10^{-4} cm^2.

Solution. If the base doping is 10^{18} cm^{-3}, the electron mobility in the base can be determined from Figure 2.8 to be $\mu_n \approx 250$ cm^{2}V^{-1}s^{-1}. Using the Einstein relationship, the diffusivity of electrons in the base is

$$D_{nB} = \frac{kT}{q}\mu_n = (0.026\text{ V})(250\text{ cm}^2\text{V}^{-1}\text{s}^{-1}) = 6.5\text{ cm}^2\text{s}^{-1}$$

and the Shockley equation for the collector current is

$$I_C \approx \frac{qAD_{nB}n_i^2}{W_B N_{AB}}\left(e^{qV_{BE}/kT}-1\right) \approx I_S\left(e^{qV_{BE}/kT}-1\right).$$

Therefore, the reverse saturation current for the transistor is

$$I_S \approx \frac{qAD_{nB}n_i^2}{W_B N_{AB}} = \frac{(1.6\times10^{-19}\text{ C})(10^{-4}\text{ cm}^2)(6.5\text{ cm}^2\text{s}^{-1})(1.45\times10^{10}\text{ cm}^{-3})^2}{(100\times10^{-7}\text{ cm})(10^{18}\text{ cm}^{-3})}$$

$$= 2.2\times10^{-15}\text{ A}.$$

4.5 SPICE Model

SPICE uses a Gummel–Poon model for the bipolar transistor. Unlike the simpler Ebers–Moll model, the Gummel–Poon model can be used for transient simulations as well as DC calculations. Figure 4.8 shows the SPICE

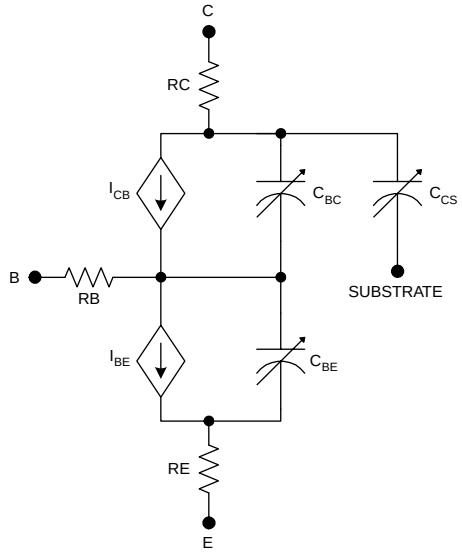


FIGURE 4.8
BJT SPICE model.

circuit model for the bipolar junction transistor. I_{BE} and I_{CB} are Shockley-type current sources including adjustable emission coefficients. These currents are calculated by

$$I_{CB} = IS\left[\exp\left(\frac{qV_{BE}}{NFkT}\right) - \exp\left(\frac{qV_{BC}}{NRkT}\right)\right]\left[1 - \frac{V_{BC}}{VAF}\right] - \frac{IS}{BR}\left[\exp\left(\frac{qV_{BC}}{NRkT}\right) - 1\right] \quad (4.33)$$

and

$$I_{BE} = IS\left[\exp\left(\frac{qV_{BE}}{NFkT}\right) - \exp\left(\frac{qV_{BC}}{NRkT}\right)\right]\left[1 - \frac{V_{BC}}{VAF}\right] + \frac{IS}{BF}\left[\exp\left(\frac{qV_{BE}}{NFkT}\right) - 1\right], \quad (4.34)$$

where

V_{BE} = base–emitter voltage
V_{BC} = base–collector voltage
IS = junction saturation current
NF = forward emission coefficient
NR = reverse emission coefficient
BF = forward beta
BR = reverse beta
VAF = forward Early voltage
ϕ_T = thermal voltage (26 mV at 300 K)

C_{BE} and C_{BC} are the base–emitter and base–collector capacitances, respectively, and include the depletion and the diffusion contributions. C_{CS} is the collector–substrate capacitance, which is a depletion layer capacitance. These capacitances are calculated by

$$C_{BE} = TF\frac{IS}{NF\phi_T}\exp\left(\frac{qV_{BE}}{NFkT}\right) + \frac{CJE}{\left(1-\frac{V_{BE}}{VJE}\right)^{MJE}}, \tag{4.35}$$

$$C_{BC} = TR\frac{IS}{NR\phi_T}\exp\left(\frac{qV_{BC}}{NRkT}\right) + \frac{CJC}{\left(1-\frac{V_{BC}}{VJC}\right)^{MJC}}, \tag{4.36}$$

and

$$C_{CS} = \frac{CJS}{\left(1-\frac{V_{CS}}{VJS}\right)^{MJS}}, \tag{4.37}$$

where

V_{BE} = base–emitter voltage
V_{BC} = base–collector voltage
IS = junction saturation current
NF = forward emission coefficient
NR = reverse emission coefficient
CJE = zero-bias base–emitter capacitance
CJC = zero-bias base–collector capacitance
CJS = zero-bias collector-substrate capacitance
VJE = base–emitter built-in potential
VJC = base–collector built-in potential
VJS = collector–substrate built-in potential
MJE = base–emitter grading coefficient
MJC = base–collector grading coefficient
MJS = collector–substrate grading coefficient

RC, *RB*, and *RE* are the series resistances in the device. The SPICE model parameters are listed in Table 4.1.

Example 4.6

Determine the value of IS for a bipolar transistor assuming that V_{BEA} = 0.7 V and that the emission coefficient is unity.

TABLE 4.1

SPICE Parameters for the BJT

Symbol	SPICE Name	Description	Units	Default	Typical
I_S	IS	Saturation current	A	1E-16	1E-16
	VAF	Forward Early voltage	V	∞	200
N_F	NF	Forward (emitter) emission coefficient	—	1.5	2
	VAR	Reverse Early voltage	V	∞	200
N_R	NR	Reverse (collector) emission coefficient	—	2	1.5
R_B	RB	Base series resistance	Ω	0	100
	RBM	Min RB at high current	Ω	RB	10
	IRB	I @ RB halfway to RBM	A	∞	0.1
R_E	RE	Emitter series resistance	Ω	0	1
R_C	RC	Collector series resistance	Ω	0	10
	CJE	Zero-bias base–emitter capacitance	F	0	2P
	VJE	Base–emitter built-in voltage	V	0.75	0.85
	MJE	Base–emitter grading coefficient	—	0.33	0.33
	CJC	Zero-bias base–collector capacitance	F	0	2P
	VJC	Base–collector built-in voltage	V	0.75	0.80
	MJC	Base–collector grading coefficient	—	0.33	0.5
	CJS	Zero-bias collector-substrate capacitance	F	0	2P
	VJS	Collector–substrate built-in voltage	V	0.75	0.80
	MJS	Collector–substrate grading coefficient	—	0.5	0.5
τ_F	TF	Forward transit time	s	0	1N
τ_R	TR	Reverse transit time	s	0	10N

Solution. Under forward active conditions,

$$I_C \approx IS \exp\left(\frac{qV_{BE}}{NFkT}\right).$$

If the collector current is 1 mA with $V_{BEA} = 0.7$ V, then

$$IS = \frac{I_C}{\exp\left(\dfrac{qV_{BE}}{NFkT}\right)} = \frac{10^{-3}\text{ A}}{\exp\left(\dfrac{0.7\text{ V}}{(1.0)(0.026\text{ V})}\right)} = 2.0\times10^{-15}\text{ A}.$$

4.6 Integrated Bipolar Junction Transistors

Integrated npn bipolar transistors are made by some variation of the double diffused epitaxial process.[13–19] In its simplest form, the process involves the growth of an epitaxial n-type collector, followed by the diffusion of a p-type base, followed by the diffusion of an n-type emitter. Many variations of this basic process exist and, increasingly, the diffusion steps are replaced by ion

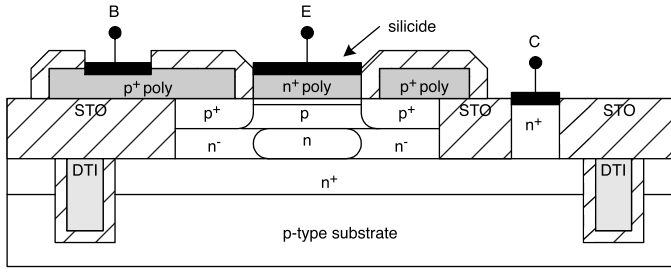


FIGURE 4.9
Integrated vertical npn bipolar transistor (STO = shallow trench oxide; DTI = deep trench isolation).

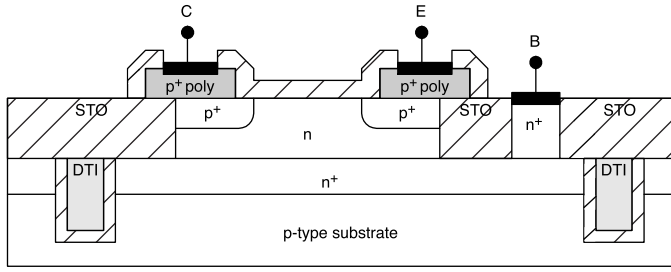


FIGURE 4.10
Integrated lateral pnp bipolar transistor (STO = shallow trench oxide; DTI = deep trench isolation).

implantation, which affords greater control. Figure 4.9 shows an integrated vertical npn bipolar transistor compatible with the process flow illustrated in Chapter 1. This device is called a vertical transistor because the current flows vertically in the base.

Pnp transistors are used rarely in digital integrated circuits; as such, the process flow is optimized for npn transistors. The design of pnp transistors usually represents a compromise between the necessary performance and the capability of fabrication with vertical npn devices. The most common type of design is a lateral pnp device like that shown in Figure 4.10. This is called a lateral transistor because the current flows laterally (parallel to the surface) in the base. Such transistors usually have poor betas because the base width is defined by lithographic processing and is relatively large.

4.7 PSPICE Simulations

Simulations were performed using PSPICE 9.1, which can be downloaded free.[20] The BJT model parameters used in all simulations are provided in Table 4.2.

TABLE 4.2

BJT SPICE Parameters

Parameter	Value	Units
IS	2.0f	A
BF	45	—
NF	1	—
BR	1	—
NR	1	—
VAF	50	V
CJE	0.7p	F
VJE	0.85	V
MJE	0.5	—
TF	0.5n	s
CJC	0.4p	F
VJC	0.8	V
MJC	0.5	—

4.7.1 Common Emitter Characteristics

The common emitter characteristics (I_C vs. V_{CE} with I_B as a parameter) for a digital bipolar junction transistor were simulated using the circuit of Figure 4.11. The BJT common emitter characteristics appear in Figure 4.12. In the forward active region, the curves are nearly flat due to the large value of the Early voltage (*VAF* = 50 V).

4.7.2 Base–Emitter Voltage for Forward Active Operation

The collector current vs. the base–emitter voltage characteristic was simulated using the circuit of Figure 4.13. The results in Figure 4.14 show that the base–emitter voltage for forward active operation is typically 0.7 V but can be as high as 0.75 V for high current densities. Also, the base–emitter voltage is nearly independent of the collector–emitter voltage.

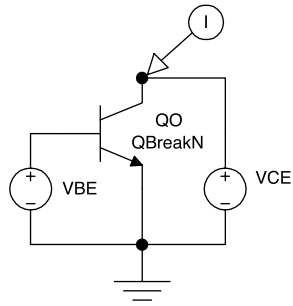


FIGURE 4.11
Circuit for simulation of the characteristic curves of a bipolar transistor.

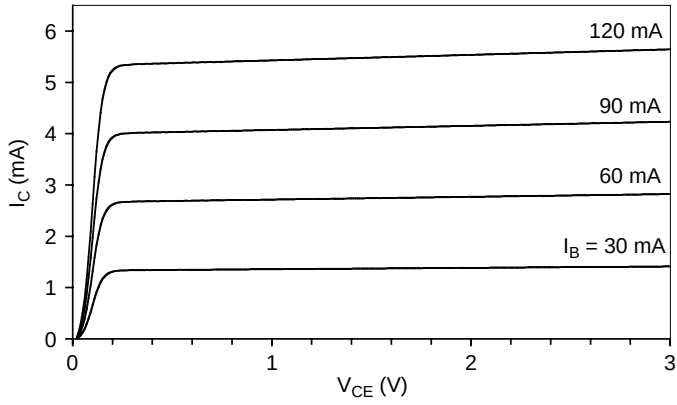


FIGURE 4.12
I_C vs. V_{CE} with I_B as a parameter.

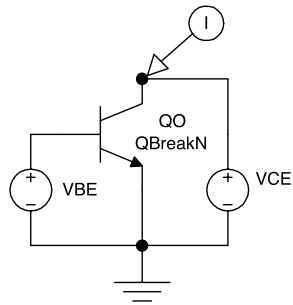


FIGURE 4.13
Circuit for simulation of the collector current vs. the base–emitter voltage for a bipolar transistor.

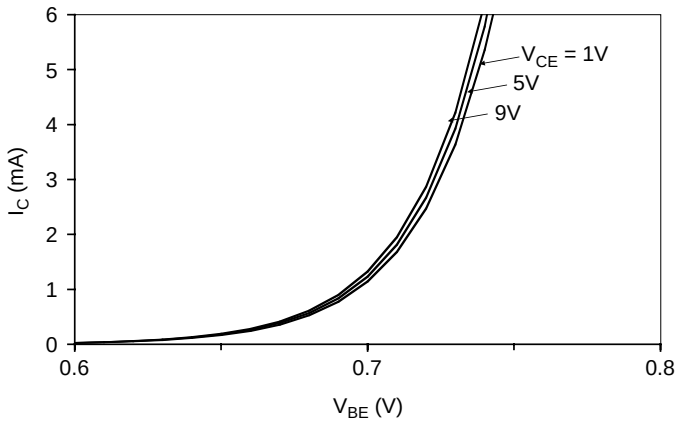


FIGURE 4.14
I_C vs. V_{BE} with V_{CE} as a parameter for a bipolar transistor.

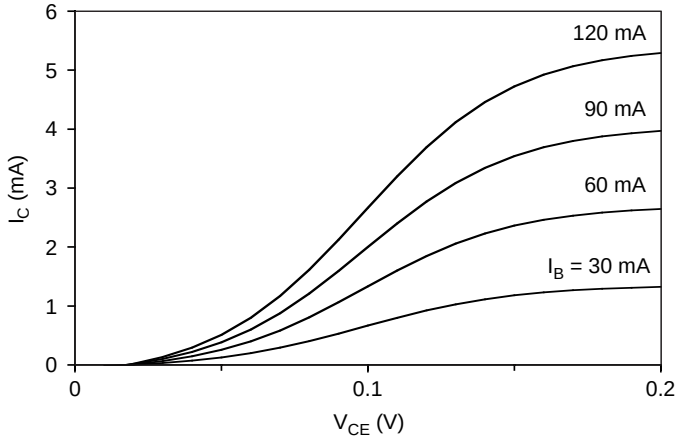


FIGURE 4.15
Simulated characteristic curves for saturation operation of a bipolar transistor.

4.7.3 Collector–Emitter Voltage for Saturation Operation

The saturation region of the characteristic curves in Figure 4.12 is expanded in Figure 4.15. It can be seen that the collector–emitter voltage varied from 0 V to approximately 0.2 V in the saturation region. It is therefore reasonable to assume V_{CES} = 0.1 V for hand calculations.

4.8 Summary

There are two basic types of bipolar junction transistors: npn and pnp; npn bipolar transistors are preferred because of their superior performance. The four DC modes of operation for a bipolar junction transistor are *forward active*, *reverse active*, *saturation*, and *cutoff*. In the forward active mode of operation, the base–emitter junction is forward biased while the base–collector junction is reverse biased. The device acts like a current-controlled current source; I_B is the controlling current and I_C is the controlled current. The common emitter current gain is β_F. For saturation operation, both junctions are forward biased and the device acts like a closed switch. In the cutoff mode, both junctions are reverse biased and the device acts like an open circuit.

In saturated bipolar digital circuits such as TTL, the transistors switch between cutoff and saturation. In current mode bipolar digital circuits such as ECL, the transistors switch between cutoff and forward active operation.

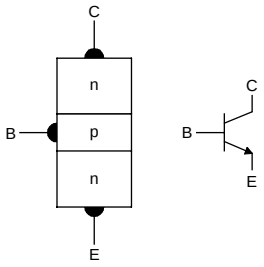
The DC operation (all four modes) of the BJT can be modeled using the Ebers–Moll model. The BJT SPICE model is more complex and includes the capacitive effects due to depletion layers and minority carrier storage.

For hand analysis of digital circuits, extremely simply models are used for the bipolar transistor. In forward active operation, it is assumed that the

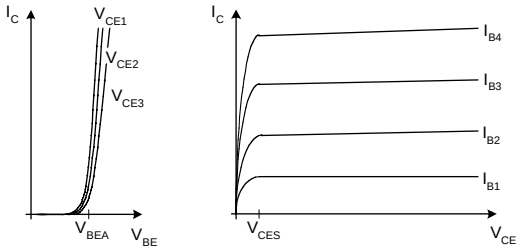
base–emitter voltage is about 7/10 of a volt. In saturation operation, it is assumed that the base–emitter voltage is about 8/10 of a volt and the collector–emitter voltage is about 1/10 of a volt.

BIPOLAR JUNCTION TRANSISTOR QUICK REFERENCE

npn BJT



npn BJT Characteristics



Bipolar junction transistors comprise two interacting p-n junctions. There are two types: npn and pnp. There are four modes of operation: cutoff, forward active, saturation, and reverse active.

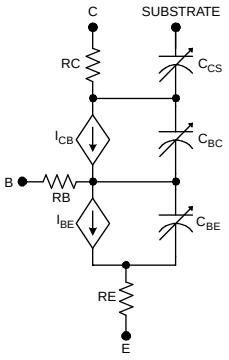
npn Modes of Operation

cutoff	$V_{BE} < 0$	$V_{BC} < 0$	$I_C \approx I_E \approx I_B \approx 0$	
forward active	$V_{BE} > 0$	$V_{BC} < 0$	$V_{BE} \approx V_{BEA} = 0.7V$	$I_C \approx \beta_F I_B$
reverse active	$V_{BE} < 0$	$V_{BC} > 0$	$V_{BC} \approx V_{BCA} = 0.7V$	
saturation	$V_{BE} > 0$	$V_{BC} > 0$	$V_{BE} \approx V_{BES} = 0.8V$	$V_{CE} \approx V_{CES} = 0.1V$

Current Gain Parameters

$$\beta_F = \frac{\alpha_F}{1-\alpha_F} \qquad \alpha_F = \frac{\beta_F}{\beta_F+1} \qquad \beta_R = \frac{\alpha_R}{1-\alpha_R} \qquad \alpha_R = \frac{\beta_R}{\beta_R+1}$$

SPICE Model



$$I_{CB} = IS\left[\exp\left(\frac{V_{BE}}{NF\phi_T}\right) - \exp\left(\frac{V_{BC}}{NR\phi_T}\right)\right] \times \left[1 - \frac{V_{BC}}{VAF}\right] - \frac{IS}{BR}\left[\exp\left(\frac{V_{BC}}{NR\phi_T}\right) - 1\right]$$

$$I_{BE} = IS\left[\exp\left(\frac{V_{BE}}{NF\phi_T}\right) - \exp\left(\frac{V_{BC}}{NR\phi_T}\right)\right] \times \left[1 - \frac{V_{BC}}{VAF}\right] + \frac{IS}{BF}\left[\exp\left(\frac{V_{BE}}{NF\phi_T}\right) - 1\right]$$

$$C_{BE} = TF\frac{IS}{NF\phi_T}\exp\left(\frac{V_{BE}}{NF\phi_T}\right) + \frac{CJE}{\left(1 - \frac{V_{BE}}{VJE}\right)^{MJE}}$$

$$C_{BC} = TR\frac{IS}{NR\phi_T}\exp\left(\frac{V_{BE}}{NR\phi_T}\right) + \frac{CJC}{\left(1 - \frac{V_{BC}}{VJC}\right)^{MJC}}$$

$$C_{CS} = \frac{CJS}{\left(1 - \frac{V_{CS}}{VJS}\right)^{MJS}}$$

Design Rules

$V_{BEA} = 0.7V$ $V_{BES} = 0.8V$ $V_{CES} = 0.1V$ $V_{BCA} = 0.7V$

f = 10^{-15} p = 10^{-12} n = 10^{-9} μ = 10^{-6} m = 10^{-3} k = 10^{3} M = 10^{6} G = 10^{9}

Laboratory Exercises

L4.1. For 2N3904 npn bipolar transistor, determine V_{BEA} @ I_C = 5 mA. Does the base–emitter voltage depend on the collector–emitter voltage? Determine V_{BES} and V_{CES} @ I_C = 5 mA, I_B = 1 mA.

L4.2. For a commercially available discrete bipolar transistor, make measurements to determine β_F and β_R. Obtain the data sheet for the transistor from the Internet and compare your beta values to those provided by the manufacturer.

L4.3. For a commercially available bipolar transistor, measure and plot the characteristic curves. Vary the collector current up to one half of the rated absolute maximum from the data sheet. Vary the collector–emitter voltage up to one half of the rated absolute maximum from the data sheet. Using the characteristic curves, plot the forward beta vs. the collector current. Compare your results to those from the data sheet. What value of collector current results in the maximum value of current gain?

L4.4. For a commercially available BJT, measure the capacitance vs. the reverse bias for the base–emitter junction. Repeat for the base–collector junction. Comment on the differences between the two junctions.

Problems

P4.1. Estimate β_F for an npn bipolar transistor assuming that the emitter injection efficiency and the collector multiplication factor are both unity. The base width is 120 nm, the base doping is 10^{18} cm^{-3}, and the base minority carrier lifetime is 1.0 ns.

P4.2. Estimate β_F for a lateral pnp bipolar transistor assuming that emitter injection efficiency and the collector multiplication factor are both unity. The base width is 250 nm, the base doping is 10^{17} cm^{-3}, and the base minority carrier lifetime is 3.0 ns.

P4.3. Estimate the base transit time for an npn bipolar transistor if the undepleted base width is 120 nm and the base doping is 10^{18} cm^{-3}.

P4.4. Estimate the base transit time for a lateral pnp bipolar transistor if the undepleted base width is 250 nm and the base doping is 10^{17} cm^{-3}.

P4.5. Estimate the reverse saturation current for an npn bipolar transistor if the undepleted base width is 120 nm, the base doping is 10^{18} cm^{-3}, the base minority carrier lifetime is 1.0 ns, and the emitter area is 10^{-4} cm^2.

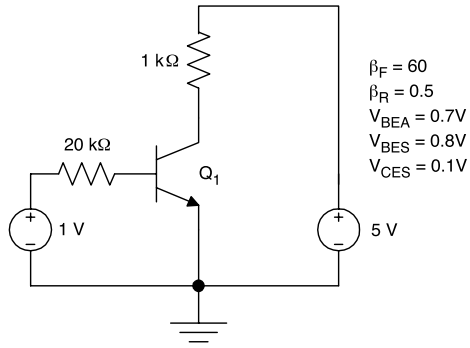


FIGURE 4.16
Circuit (P4.7).

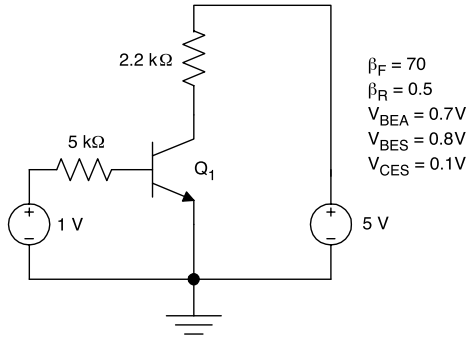


FIGURE 4.17
Circuit (P4.8).

P4.6. Estimate the emitter area for an npn bipolar transistor if the forward active collector current is 1 mA with a base–emitter voltage of 0.75 V. The base width is 180 nm and the base doping is 10^{17} cm^{-3}. Assume the forward emission coefficient is unity and the series resistances are negligible.

P4.7. For the bipolar transistor in the circuit of Figure 4.16, determine the mode of operation and the collector–emitter voltage.

P4.8. For the bipolar transistor in the circuit of Figure 4.17, determine the collector current.

P4.9. Determine the current I_Z in the circuit of Figure 4.18.

P4.10. Using SPICE, determine and plot I_C vs. V_{CE} with I_B as a parameter for a bipolar transistor with: IS = 1 nA, BF = 70, NF = 1, BR = 1, NR = 1, and VAF = 25. Use a nested sweep to obtain a single plot and consider the cases I_B = 0, 25, 50, 75, and 100 μA.

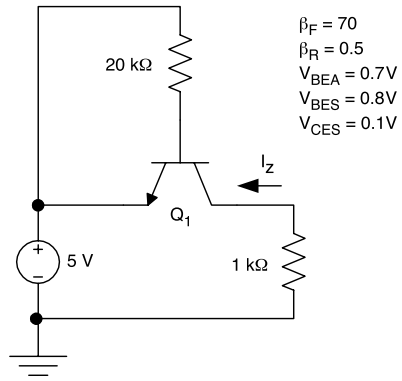


FIGURE 4.18
Circuit (P4.9).

References

1. Ghandhi, S.K., *The Theory and Practice of Microelectronics*, Krieger, Malabar, FL, 1968.
2. Grove, A.S., *Physics and Technology of Semiconductor Devices*, John Wiley & Sons, New York, 1967.
3. Streetman, B.G. and Banerjee, S., *Solid State Electronic Devices*, Prentice Hall, Englewood Cliffs, NJ, 1999.
4. Sze, S.M., *Semiconductor Devices: Physics and Technology*, 2nd ed., John Wiley & Sons, New York, 2001.
5. Sze, S.M., *Physics of Semiconductor Devices*, John Wiley & Sons, New York, 1981.
6. Taur, Y. and Ning, T.H., *Fundamentals of Modern VLSI Devices*, Cambridge University Press, New York, 1998.
7. Ashburn, P., *Design and Realization of Bipolar Transistors*, John Wiley & Sons, New York, 1988.
8. Ning, T.H., History and future perspective of the modern silicon bipolar transistor, *IEEE Trans. Electron. Devices*, 48, 2485, 2001.
9. Asbeck, P.M. and Nakamura, T., Bipolar transistor technology: past and future directions, *IEEE Trans. Electron. Devices*, 48, 2455, 2001.
10. Kapoor, A.K. and Roulston, D.J., Eds., *Polysilicon Emitter Bipolar Transistors*, IEEE Press, New York, 1989.
11. Shockley, W., The theory of p–n junctions in semiconductors and p–n junction transistors, *Bell Sys. Tech. J.*, 28, 435, 1949.
12. Ebers, J.J. and Moll, J.L., Large-signal behavior of junction transistors, *Proc. IRE*, 42, 1761, 1954.
13. Ghandhi, S.K., *VLSI Fabrication Principles*, 2nd ed., John Wiley & Sons, New York, 1992.
14. Shiba, T., Uchino, T., Ohnishi, K., and Tamaki, Y., *In-situ* phosphorus-doped polysilicon emitter technology for very high-speed, small emitter bipolar transistors, *IEEE Trans. Electron. Devices*, 43, 889, 1996.

15. Warnock, J.D., Silicon bipolar device structures for digital applications: technology trends and future directions, *IEEE Trans. Electron. Devices*, 42, 377, 1995.
16. Yoshino, C., Inou, K., Matsuda, S., Nakajima, H., Tsuboi, Y., Naruse, H., Sugaya, H., Katsumata, Y., and Iwai, H., A 62.8-GHz f_{max} LP-CVD epitaxially grown silicon-base bipolar transistor with extremely high Early voltage of 85.7 V, *1995 Symp. VLSI Technol. Dig. Tech. Papers*, 131, 1995.
17. Chen, T.C., Cressler, J.D., Toh, K.Y., Warnock, J., Lu, P.F., Jenkins, K.A., Basavaiah, S., Manny, M.P., Ng, H.Y., Tang, D.D., Li, G.P., Chuang, C.T., Polcari, M.R., Ketchen, M.B., and Ning, T.H., A submicron high-performance bipolar technology, *1989 Symp. VLSI Technol. Dig. Tech. Papers*, 87, 1989.
18. Ning, T.H., Isaac, R.D., Solomon, P.M., Tang, D.D., Yu, H.N., Feth, G.C., and Wiedmann, S.K., Self-aligned bipolar transistors for high-performance and low-power-delay VLSI, *IEEE Trans. Electron. Devices*, 28, 1010, 1981.
19. Tang, D.D. and Solomon, P.M., Bipolar transistor design for optimized power-delay logic circuits, *IEEE J. Solid-State Circuits*, 14, 679, 1979.
20. www.cadence.com.

5

Transistor–Transistor Logic

5.1 Introduction

Transistor–transistor logic (TTL) evolved from resistor–transistor logic (RTL) and diode–transistor logic (DTL). All of these circuit families use bipolar transistors as saturated switches and are collectively called *saturated bipolar logic circuits*. The delay associated with bringing transistor switches out of saturation is an important speed limitation in all such circuits. However, TTL uses active pull-up circuitry and avoids the long RC delays present in RTL and DTL, which use pull-up resistors.

Numerous circuit refinements have brought further advances in switching speed and have maintained the commercial importance of TTL. Today high-speed Schottky TTL circuitry is used extensively in applications requiring high off-chip data rates with highly capacitive loads. These applications include the motherboards of desktop computers and servers.

5.2 Circuit Evolution

The simplest bipolar logic gate comprises a single bipolar transistor, used as a switch, and two resistors, as shown in Figure 5.1. The transistor normally operates in cutoff (off state) or saturation (on state). This type of logic circuit is called resistor–transistor logic (RTL). A basic limitation of the RTL circuit is the unfavorable trade-off between the fan-out and logic swing. With a logic-zero output, Q_O is saturated and the output low voltage is

$$V_{OL} = V_{CES}, \tag{5.1}$$

regardless of the fan-out. With a logic-one output, Q_O is cut off. With no collector current flowing, there is no voltage drop in the collector resistor and the output voltage is equal to the supply voltage:

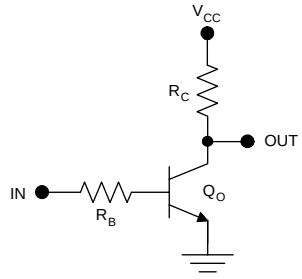


FIGURE 5.1
RTL inverter.

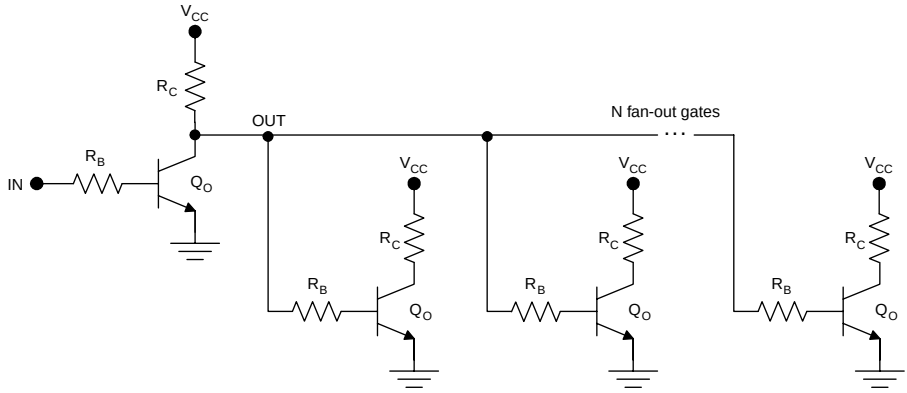


FIGURE 5.2
RTL inverter loaded by N identical fan-out gates.

$$V_{OH} = V_{CC}. \tag{5.2}$$

The connection of fan-out gates alters V_{OH} by introducing a voltage drop across R_C.

Suppose that N fan-out gates are connected to the output of a driving gate as shown in Figure 5.2. Following the usual approach, it is assumed that the fan-out gates are identical in design to the driving gate. This situation is most easily analyzed using the Thevenin equivalent for the fan-out gates. To determine the Thevenin equivalent for the load gates, the transistor in each load gate is replaced with a voltage source equal to V_{BES}, as shown in Figure 5.3. After substituting the Thevenin equivalent of the N fan-out gates, the circuit model shown in Figure 5.4 is obtained and the output circuitry of the load gates is ignored for this calculation. Using the simplified model of Figure 5.4, V_{OH} can be determined using the voltage divider rule and the assumption that Q_O is cut off. Thus

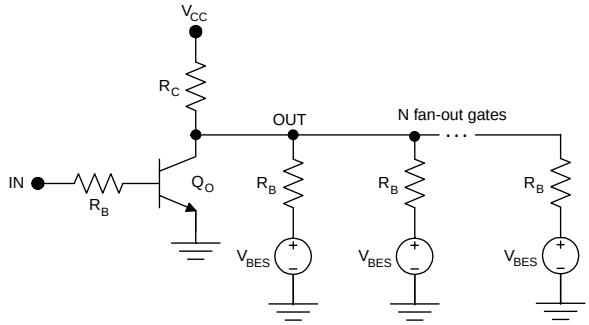


FIGURE 5.3
RTL inverter with N fan-out gates; each gate is modeled as an ideal voltage source in series with a resistor.

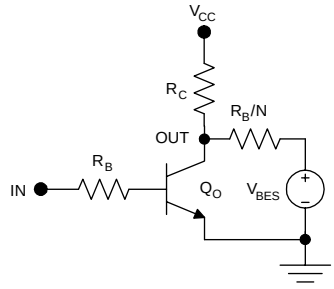


FIGURE 5.4
RTL inverter with Thevenin equivalent of N fan-out gates.

$$V_{OH} = V_{BES} + \frac{\left(V_{CC} - V_{BES}\right) R_B/N}{R_C + R_B/N}, \tag{5.3}$$

where N is the number of fan-out gates. Therefore, the logic swing of the RTL gate is given by

$$LS = V_{OH} - V_{OL} = V_{BES} + \frac{\left(V_{CC} - V_{BES}\right) R_B/N}{R_C + R_B/N} - V_{CES}, \tag{5.4}$$

which decreases with the fan-out N. Therefore, V_{OH}, the high noise margin, and the logic swing degrade with increasing fan-out in the case of RTL circuitry.

Example 5.1

Determine the logic swing as a function of the fan-out for the RTL inverter shown in Figure 5.5.

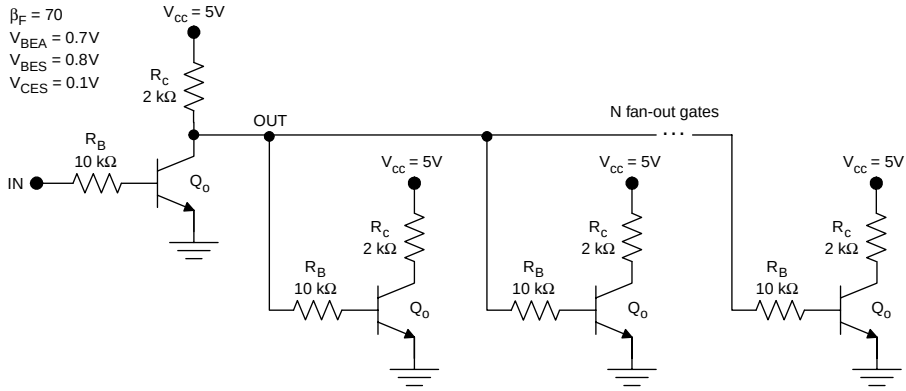


FIGURE 5.5
Loaded RTL inverter for the calculation of the logic swing.

Solution. For this circuit, the logic swing is given by

$$LS = V_{BES} + \frac{(V_{CC} - V_{BES})R_B/N}{R_C + R_B/N} - V_{CES}$$

$$= 0.8\text{ V} + \frac{(5\text{ V} - 0.8\text{ V})10\text{ k}\Omega/\text{N}}{2\text{ k}\Omega + 10\text{ k}\Omega/\text{N}} - 0.1\text{ V}$$

$$= 0.7\text{ V} + 4.2\text{ V}\left(\frac{5}{5+N}\right)$$

The results are plotted in Figure 5.6.

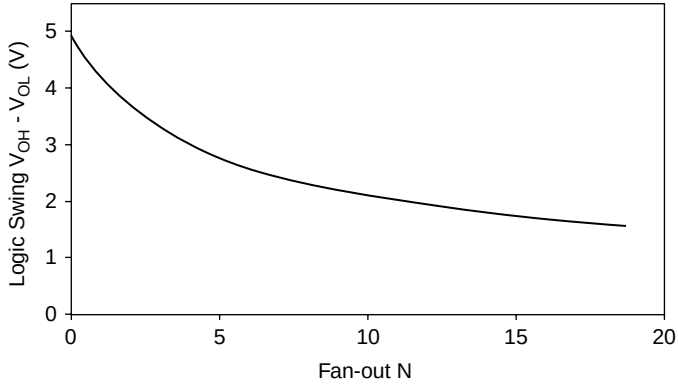


FIGURE 5.6
Logic swing vs. fan-out N for the RTL inverter.

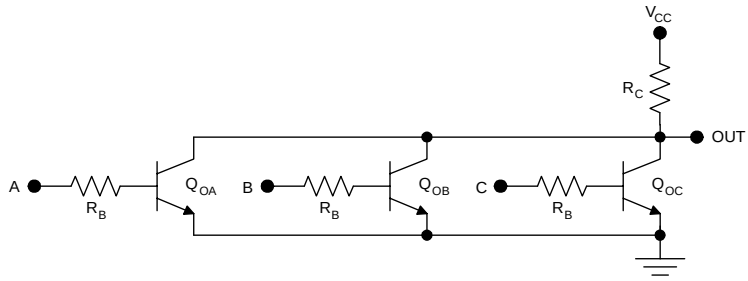


FIGURE 5.7
RTL NOR3 gate.

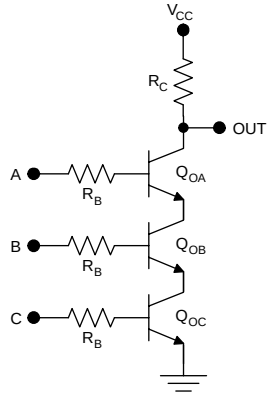


FIGURE 5.8
RTL NAND3 gate.

Arbitrary logic functions may be realized in RTL by the parallel or series connection of the switch transistors. For example, the three-input RTL NOR gate is constructed as shown in Figure 5.7. Here, the application of logic one to any of the inputs causes the associated transistor to saturate, bringing the output low.

RTL NAND gates are made by placing the switch transistors in series as shown in Figure 5.8. Here, the output low voltage degrades with the fan-in M (the number of inputs) because

$$V_{OL} = MV_{CES}. \tag{5.5}$$

Another difficulty with RTL NAND gates is that V_{IH} is different for each input. For the A input (V_{INA}), the value of V_{IH} is the same as for the inverter. However, for the B input (V_{INB}), the value of V_{IH} is larger by V_{CES}. For the C input, the value of V_{IH} is increased by $2V_{CES}$ compared to the A input. The topmost input exhibits the largest value of V_{IH}; therefore, it exhibits a degraded logic-one noise margin compared to the inverter.

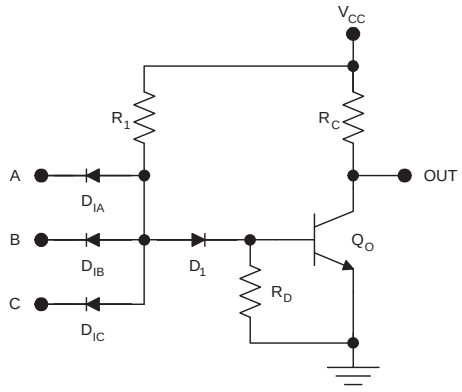


FIGURE 5.9
Elementary DTL NAND3 gate with one bipolar transistor.

Logic design with RTL is quite restrictive because the logic swing and noise margins depend on the fan-out and fan-in (for NAND gates). Diode–transistor logic (DTL) evolved from RTL as a solution to this problem. In the basic DTL gate of Figure 5.9, the diodes prevent loading of the driving gate under output high conditions. Therefore, the logic swing does not degrade with the fan-out. Furthermore, the NAND function is obtained without placing switch transistors in series, so the input levels are independent of the fan-in.

Transistor–transistor logic (TTL) evolved from this type of elementary DTL-type circuit. Consider the DTL inverter shown in Figure 5.10. The simplest form of TTL can be realized by replacing the back-to-back diodes with an npn bipolar transistor. This amounts to merging the p-regions of the diodes together as the base of the transistor. The resulting basic TTL circuit is shown in Figure 5.11. The merging of the back-to-back diodes into a single bipolar transistor results in a more compact structure; thus, the circuit consumes less chip area and has reduced parasitic capacitances. The reduction in parasitic capacitances is accompanied by reduced RC time constants and improved dynamic response. A second advantage of the TTL circuit is that the input transistor becomes forward active during high-to-low transitions at the input. This transistor action, which is absent in the DTL version, speeds the removal of base charge from Q_O. The immediate benefit is a shortening of the saturation delay and the low-to-high propagation delay.

As a further improvement, the addition of an emitter follower as shown in Figure 5.12 greatly increases the available base drive for the output transistor. Two important benefits stem from this modification. First, the improved DC base drive for Q_O improves the DC fan-out and, second, the improved base drive for Q_O during input low-to-high transitions improves t_{PHL}. Standard TTL circuitry embodies two more improvements as shown in Figure 5.13. The first is the addition of active pull-up circuitry; the second is the inclusion of antiringing circuitry in the form of D_1.

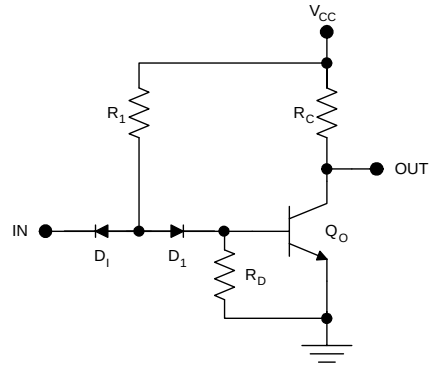


FIGURE 5.10
DTL-type inverter circuit.

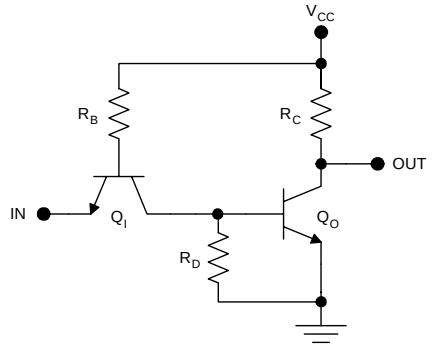


FIGURE 5.11
Basic TTL inverter circuit.

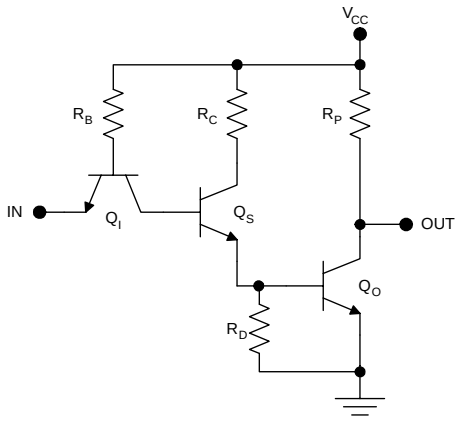


FIGURE 5.12
Improved TTL inverter.

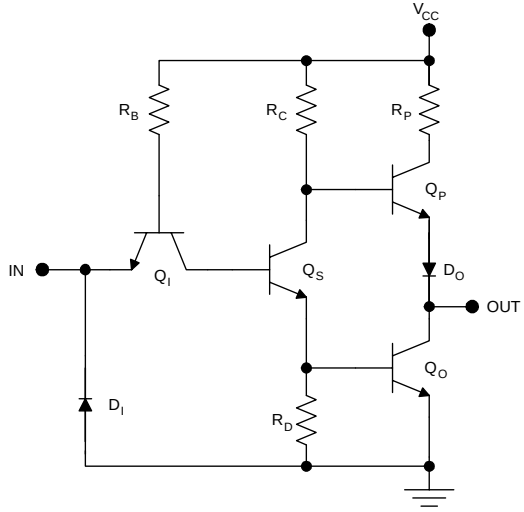


FIGURE 5.13
Standard TTL inverter.

The active pull-up circuitry embodied in Q_P greatly improves t_{PLH}. Without active pull-up, the choice of R_C involves a direct trade-off between P_L and t_{PLH}. The low-to-high propagation delay is limited by the $R_C C_L$ time constant, where C_L is the load capacitance. Therefore, improvement in t_{PLH} requires a reduction in R_C. This, in turn, increases P_L. It is thus advantageous to uncouple these two design issues so that each can be optimized.

Under steady DC conditions with a low output, Q_P is cut off, so the dissipation is minimized regardless of the value of R_P. During a low-to-high transition at the output, Q_P becomes forward active and acts as an emitter follower. The resulting low output impedance gives rise to a low RC time constant and a rapid low-to-high transition at the output.

It is necessary to include a series diode D_O in the pull-up circuitry to ensure that Q_P can turn off when the output goes low. This can be shown as follows. With the output steady at its logic-zero level, Q_S and Q_O are saturated. Under these conditions, the voltage at the base of Q_P is

$$V_{BP} = V_{BES} + V_{CES} \tag{5.6}$$

and the voltage at the emitter of Q_P is

$$V_{EP} = V_{CES} + V_D . \tag{5.7}$$

Therefore, the base-emitter voltage for Q_P is

$$V_{BEP} = V_{BP} - V_{EP} = V_{BES} - V_D . \tag{5.8}$$

This value is typically 0.1 V, so Q_P will be safely cut off with logic-zero output, as desired.

If the diode is omitted, then the pull-up transistor cannot turn off. Under logic-zero output conditions, with no series diode in the output circuitry, the voltage at the base of Q_P is unchanged. However, the voltage at the emitter of Q_P with no series diode is

$$V_{EP} = V_{CES} . \tag{5.9}$$

Therefore, the base-emitter voltage for Q_P with the series diode omitted is

$$V_{BEP} = V_{BP} - V_{EP} = V_{BES} ; \tag{5.10}$$

thus, the pull-up transistor is unable to turn off if the series diode is omitted.

The right-hand side of the standard TTL circuit is often called a "totem pole output" because of the appearance of the stacked components. The transistor Q_S is called the "drive splitter" because Q_O is driven by the emitter of the drive splitter, resulting in noninverting emitter follower action. On the other hand, Q_P is driven by the collector of the drive splitter, which results in inverting common emitter action. Therefore, Q_S drives Q_O or Q_P, but not both. It is thus said that Q_S "splits the drive" between the two output transistors.

5.3 Using Kirchhoff's Voltage Law (KVL) in TTL Circuits

The currents in a BJT circuit such as a TTL inverter can be determined using Ohm's law and Kirchhoff's voltage law (KVL). According to KVL, "the sum of the voltages around a loop is 0." Consider the TTL inverter shown in Figure 5.14 and assume that the input voltage is such that the input transistor is saturated. In order to determine the base current for the input transistor, KVL can be applied to the highlighted loop. Proceeding counterclockwise around the loop,

$$V_{IN} - V_{CC} + V_{RB} + V_{BEI} = 0 . \tag{5.11}$$

The terms are added if the plus sign is encountered first (moving counterclockwise) and subtracted if the minus sign is encountered first. Rearranging yields

$$V_{RB} = V_{CC} - V_{BEI} - V_{IN} . \tag{5.12}$$

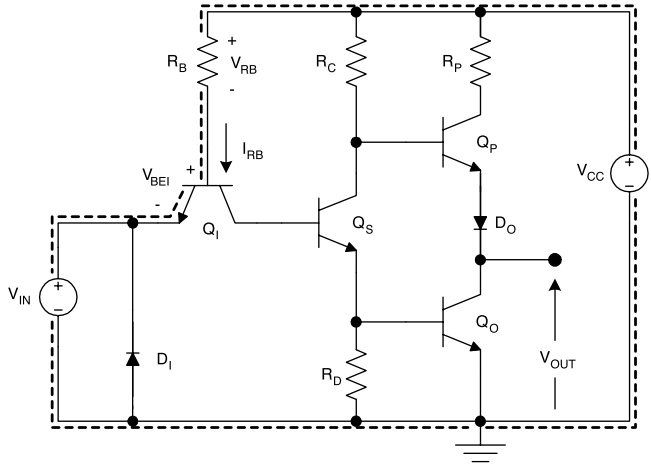


FIGURE 5.14
TTL inverter circuit for the determination of the base current in Q_I with the input transistor saturated.

Then, using Ohm's law, the current in R_B (the base current) can be determined:

$$I_{RB} = \frac{V_{RB}}{R_B} = \frac{V_{CC} - V_{BEI} - V_{IN}}{R_B}. \tag{5.13}$$

On the other hand, the base current may be determined in one step using the *KVL shortcut method*:

$$I_R = \frac{\sum\begin{pmatrix}\text{source}\\ \text{voltages}\end{pmatrix} - \sum\begin{pmatrix}\text{voltage}\\ \text{drops}\end{pmatrix}}{R} = \frac{V_{CC} - V_{BEI} - V_{IN}}{R_B}. \tag{5.14}$$

In the application of this method, the path must be chosen so that only the voltage drop to be determined is unknown. (The voltages across reverse-biased p–n junctions are unknown until determined based on other circuit constraints.)

As another example, suppose the input voltage is sufficiently positive to cause Q_S and Q_O to saturate. Then the KVL analysis can be applied to the loop highlighted in Figure 5.15. Using the KVL shortcut method for this situation yields

$$I_R = \frac{\sum\begin{pmatrix}\text{source}\\ \text{voltages}\end{pmatrix} - \sum\begin{pmatrix}\text{voltage}\\ \text{drops}\end{pmatrix}}{R} = \frac{V_{CC} - V_{BCI} - V_{BES} - V_{BEO}}{R_B}. \tag{5.15}$$

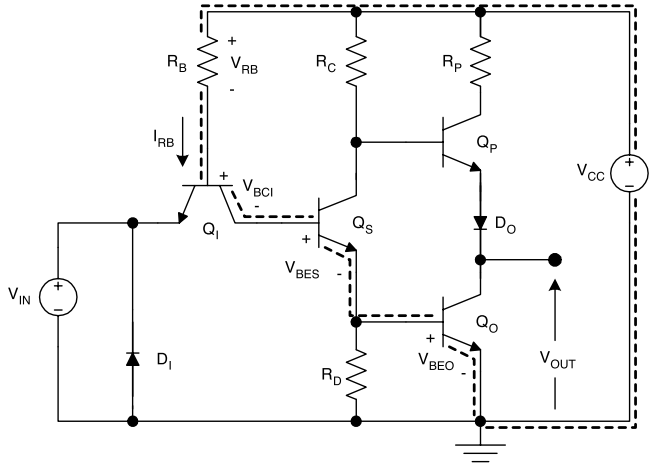


FIGURE 5.15
TTL inverter circuit for the determination of the base current in Q_I with Q_S and Q_O saturated.

This example illustrates that the choice of the loop for application of KVL depends on the modes of operation for the transistors.

The KVL shortcut method greatly simplifies the analysis of circuits involving many transistors and will be used throughout this book, as will simplified circuit diagrams. Therefore, the loops used for KVL analysis will not always appear explicitly. In addition, the polarities of voltages and currents will normally be omitted for simplicity's sake unless their omission results in an ambiguity.

5.4 Voltage Transfer Characteristic

To determine the voltage transfer characteristic for the standard TTL circuit shown in Figure 5.16, imagine that the input voltage is gradually increased from 0. With $V_{IN} = 0$, Q_I is saturated. There is no base drive for Q_S, so Q_S and Q_O are cut off; Q_P is forward active. If the small voltage drop in R_C is neglected (due to the base current of the pull-up transistor), then the output voltage is

$$V_{OH} = V_{CC} - V_{BEA} - V_D . \tag{5.16}$$

The first breakpoint in the VTC occurs at the input voltage for which Q_S turns on. Once Q_S becomes forward active, this will create a significant voltage drop in R_C that will, in turn, reduce the output voltage. Therefore, V_{IL} is the input voltage at the first breakpoint, given by

$$V_{IL} = V_{BEA} - V_{CES} . \tag{5.17}$$

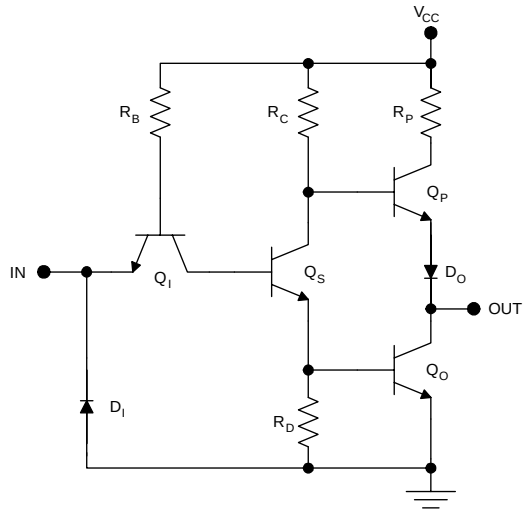


FIGURE 5.16
Standard TTL inverter.

The second breakpoint occurs at the input voltage for which Q_O turns on and occurs when the emitter current in Q_S is enough to produce a 0.7 V drop in R_D. Therefore, the input and output voltages corresponding to the second breakpoint are

$$V_{\text{IN2}} = 2V_{BEA} - V_{CES} \tag{5.18}$$

and

$$V_{\text{OUT2}} = V_{CC} - R_C\left(\frac{V_{BEA}}{R_D}\right) - V_{BEA} - V_D\,. \tag{5.19}$$

The third breakpoint occurs at the input voltage for which Q_S and Q_O saturate. For the purpose of approximate calculations, assume that the two transistors saturate together. The input and output voltages corresponding to the third breakpoint are

$$V_{IH} = 2V_{BES} - V_{CES} \tag{5.20}$$

and

$$V_{OL} = V_{CES}\,. \tag{5.21}$$

At this breakpoint, the input transistor is still saturated; however, if the input voltage is raised sufficiently high, Q_I will become reverse active. This does not affect the output voltage, which remains at V_{CES}. Therefore, the change in the mode of operation for Q_I cannot be seen in the voltage transfer characteristic.

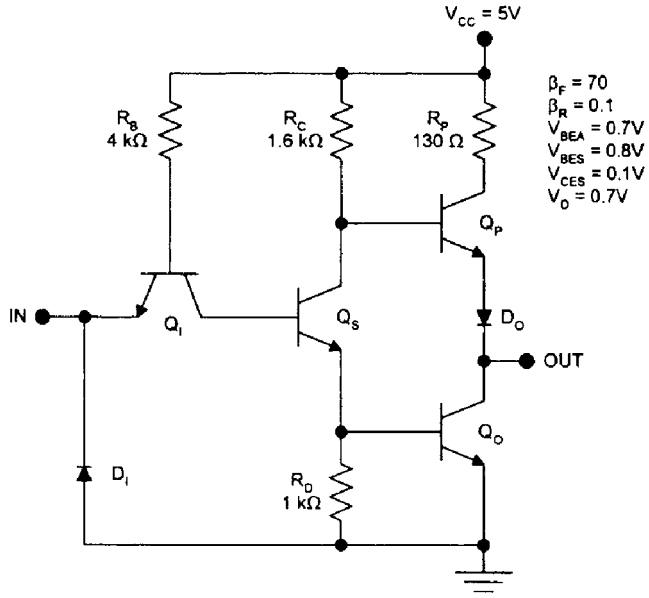


FIGURE 5.17
Example TTL inverter for the calculation of the VTC.

Example 5.2

Determine the voltage transfer characteristic for the standard TTL inverter of Figure 5.17.

Solution. At the first breakpoint in the characteristic,

$$V_{IL} = V_{BEA} - V_{CES} = 0.7\ \text{V} - 0.1\ \text{V} = 0.6\ \text{V}$$

and

$$V_{OH} = V_{CC} - V_{BEA} - V_D = 5\ \text{V} - 0.7\ \text{V} - 0.7\ \text{V} = 3.6\ \text{V}\ .$$

At the second breakpoint,

$$V_{\text{IN2}} = 2V_{BEA} - V_{CES} = 1.4\ \text{V} - 0.1\ \text{V} = 1.3\ \text{V}$$

and

$$V_{\text{OUT2}} = V_{CC} - R_C\left(\frac{V_{BEA}}{R_D}\right) - V_{BEA} - V_D$$

$$= 5\ \text{V} - 1.6\ \text{k}\Omega\left(\frac{0.7\ \text{V}}{1\ \text{k}\Omega}\right) - 0.7\ \text{V} - 0.7\ \text{V} = 2.5\ \text{V}$$

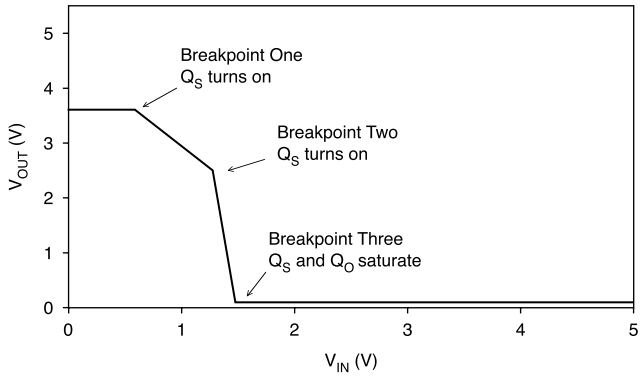


FIGURE 5.18
Voltage transfer characteristic for standard TTL inverter.

At the third breakpoint in the voltage transfer characteristic,

$$V_{IH} = 2V_{BES} - V_{CES} = 1.6\text{ V} - 0.1\text{ V} = 1.5\text{ V}$$

and

$$V_{OL} = V_{CES} = 0.1\text{ V}.$$

If the input voltage is further increased, the input transistor will become reverse active. With Q_S and Q_O saturated, the voltage at the collector of Q_I is

$$V_{CI} = 2V_{BES} = 1.6\text{ V}.$$

If the base-collector voltage is assumed to be 0.7 V, then the voltage at the base of Q_I is

$$V_{BI} = V_{CI} + V_{BCI} = 1.6\text{ V} + 0.7\text{ V} = 2.3\text{ V}.$$

The input transistor becomes reverse active if the base–emitter junction is reverse biased while the base–collector junction is forward biased. This occurs when the input voltage exceeds 2.3 V — the voltage at the base of Q_I. However, the change of mode in Q_I does not affect the output voltage, which remains at 0.1 V. The entire voltage transfer characteristic is shown in Figure 5.18.

5.5 Dissipation

The dissipation of TTL is usually dominated by the DC term, which will be considered here. For the calculation of P_L, it is customary to assume that the input voltage is a logic one level from a similar gate as shown in Figure 5.19.

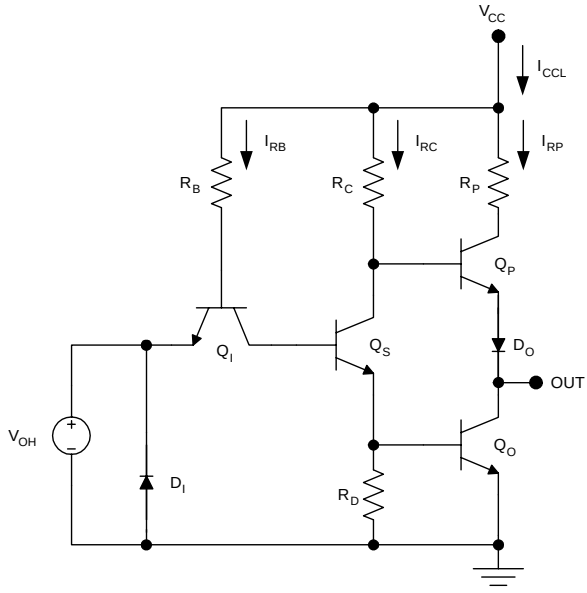


FIGURE 5.19
Standard TTL inverter for the calculation of P_L.

In this situation, Q_I is reverse active, Q_S and Q_O are saturated, and Q_P is cut off. Thus the DC dissipation with a low output is

$$\begin{aligned} P_L &= V_{CC} I_{CCL} = V_{CC}\left(I_{RB} + I_{RC} + I_{RP}\right) \\ &= V_{CC}\left(\frac{V_{CC} - V_{BCI} - V_{BESS} - V_{BESO}}{R_B} + \frac{V_{CC} - V_{CESS} - V_{BESO}}{R_C} + 0\right) \\ &= V_{CC}\left(\frac{V_{CC} - V_{BCA} - 2V_{BES}}{R_B} + \frac{V_{CC} - V_{CES} - V_{BES}}{R_C}\right) \end{aligned} \tag{5.22}$$

Current in R_P is 0 because the pull-up transistor is cut off with a low output. The exact value of V_{IN} does not appear explicitly in the equations because Q_I is reverse active. The value of V_{BCA} is usually approximated as V_{BEA}.

For the calculation of P_H, assume that the input is tied to a voltage equal to V_{OL} from a similar gate as shown in Figure 5.20. The DC dissipation with a high output is

$$\begin{aligned} P_H &= V_{CC} I_{CCH} = V_{CC}\left(I_{RB} + I_{RC} + I_{RP}\right) \\ &= V_{CC}\left(\frac{V_{CC} - V_{BESI} - V_{OH}}{R_B} + NI_{IH}\right) = V_{CC}\left(\frac{V_{CC} - V_{BES} - V_{CES}}{R_B} + NI_{IH}\right) \end{aligned} \tag{5.23}$$

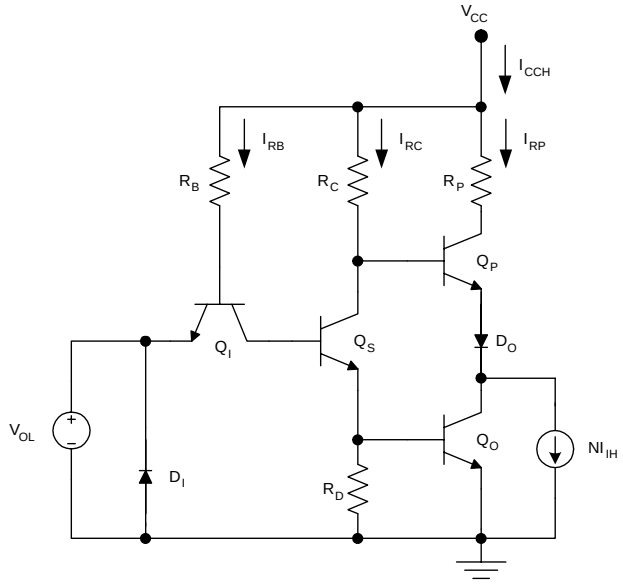


FIGURE 5.20
Standard TTL inverter for the calculation of P_H.

where N is the number of fan-out gates, each of which draws a current equal to the high input current I_{IH}. Notice that

$$I_{RC} + I_{RP} = NI_{IH} \tag{5.24}$$

because Q_S is cut off.

The input high current may be calculated based on Figure 5.21. With a logic-one input, Q_I is reverse active, Q_S and Q_O are saturated, and Q_P is cut off. Therefore, the input transistor base current is

$$I_{BI} = \frac{V_{CC} - V_{BCA} - 2V_{BES}}{R_B} \tag{5.25}$$

and

$$I_{IH} = I_{EI} = \beta_R \left(\frac{V_{CC} - V_{BCA} - 2V_{BES}}{R_B} \right). \tag{5.26}$$

Now this result can be used to determine the output high power dissipation. With N fan-out gates connected,

$$P_H = V_{CC} \left(\frac{V_{CC} - V_{BES} - V_{CES}}{R_B} + N\beta_R \left(\frac{V_{CC} - V_{BCA} - 2V_{BES}}{R_B} \right) \right). \tag{5.27}$$

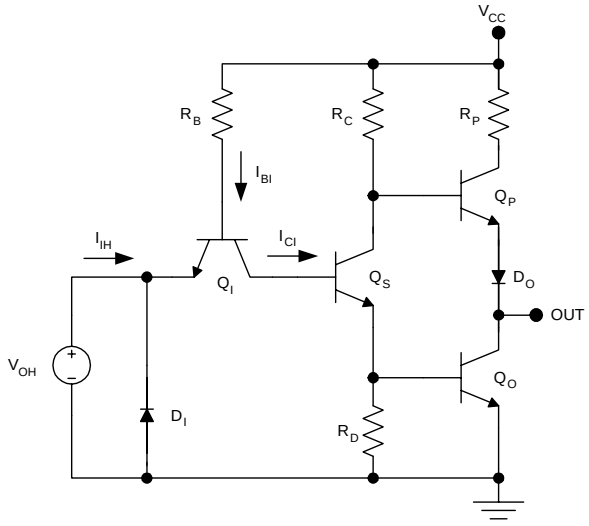


FIGURE 5.21
Standard TTL inverter for the calculation of I_{IH}.

The dissipation of standard TTL circuits can be reduced by scaling up the resistances or by reducing the supply voltage. As expected, either approach involves a trade-off between speed and power. In modern TTL gates, the power delay product is reduced by a combination of circuit and device design improvements.

Example 5.3

Calculate the average DC dissipation for the standard TTL inverter shown in Figure 5.22, assuming $N = 10$.

Solution. The output low power is

$$P_L = V_{CC}\left(\frac{V_{CC} - V_{BCA} - 2V_{BES}}{R_B} + \frac{V_{CC} - V_{CES} - V_{BES}}{R_C}\right)$$

$$= 5\text{ V}\left(\frac{5\text{ V} - 0.7\text{ V} - 1.6\text{ V}}{4\text{ k}\Omega} + \frac{5\text{ V} - 0.1\text{ V} - 0.8\text{ V}}{1.6\text{ k}\Omega}\right) = 16.2\text{ mW}.$$

The output high power is

$$P_H = V_{CC}\left(\frac{V_{CC} - V_{BES} - V_{CES}}{R_B} + N\beta_R\left(\frac{V_{CC} - V_{BCA} - 2V_{BES}}{R_B}\right)\right)$$

$$= 5\text{ V}\left[\frac{5\text{ V} - 0.8\text{ V} - 0.1\text{ V}}{4\text{ k}\Omega} + (10)(0.1)\left(\frac{5\text{ V} - 0.7\text{ V} - 1.6\text{ V}}{4\text{ k}\Omega}\right)\right] = 8.5\text{ mW}.$$

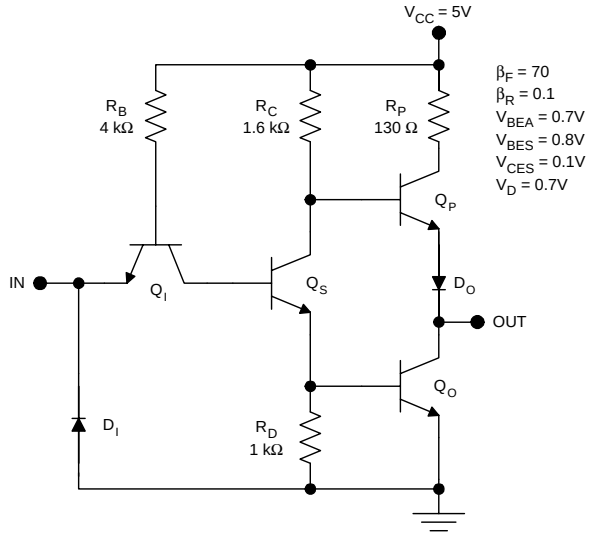


FIGURE 5.22
Standard TTL inverter for the calculation of P_{DC}.

Assuming a 50% duty cycle at the output, the average DC power is

$$P_{DC} = \frac{P_L + P_H}{2} = \frac{16.2\ \text{mW} + 8.5\ \text{mW}}{2} = 12.4\ \text{mW}.$$

5.6 Fan-Out

A DC limitation is placed on the maximum fan-out for TTL gates based on the need to maintain the output transistor in saturation with a low output. However, sometimes speed is the overriding factor in determining the maximum fan-out. Discussion is confined here to the DC limitation on the fan-out.

With a low output, the base current to the output transistor can be determined using Figure 5.23:

$$\begin{aligned} I_{BO} &= I_{ES} - I_{RD} = I_{CI} + I_{CS} - I_{RD} \\ &= (\beta_R + 1)\left(\frac{V_{CC} - V_{BCA} - 2V_{BES}}{R_B}\right) + \frac{V_{CC} - V_{CES} - V_{BES}}{R_C} - \frac{V_{BES}}{R_D} \end{aligned} \tag{5.28}$$

To keep Q_O safely in saturation and avoid a degradation of V_{OL}, it is necessary that

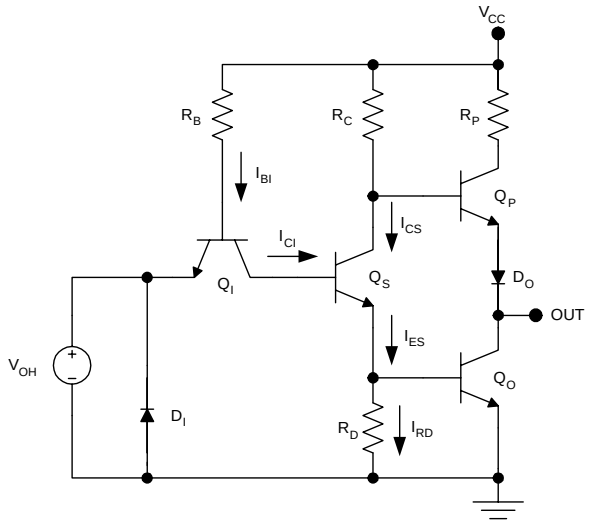


FIGURE 5.23
Standard TTL inverter for the calculation of I_{BO} and N_{MAX}.

$$I_{CO} \le \sigma_{MAX}\beta_F I_{BO}\,, \tag{5.29}$$

where σ_{MAX} is a safety factor less than one. The maximum current that the output of the gate can sink is

$$\begin{aligned} I_{OL} &= \sigma_{MAX}\beta_F I_{BO} \\ &= \sigma_{MAX}\beta_F\left[(\beta_R+1)\left(\frac{V_{CC}-V_{BCA}-2V_{BES}}{R_B}\right)+\frac{V_{CC}-V_{CES}-V_{BES}}{R_C}-\frac{V_{BES}}{R_D}\right] \end{aligned} \tag{5.30}$$

The load current contributed by each fan-out gate is I_{IL}; this can be determined with the aid of Figure 5.24, which represents the situation for the load gates. In the load gates, the input transistor is saturated but Q_S and Q_O are cut off. Therefore, for the load gates,

$$I_{IL} = \frac{V_{CC}-V_{BES}-V_{CES}}{R_B}\,. \tag{5.31}$$

The maximum fan-out based on DC loading is the largest integer satisfying

$$N_{MAX} \le \frac{I_{OL}}{I_{IL}}\,. \tag{5.32}$$

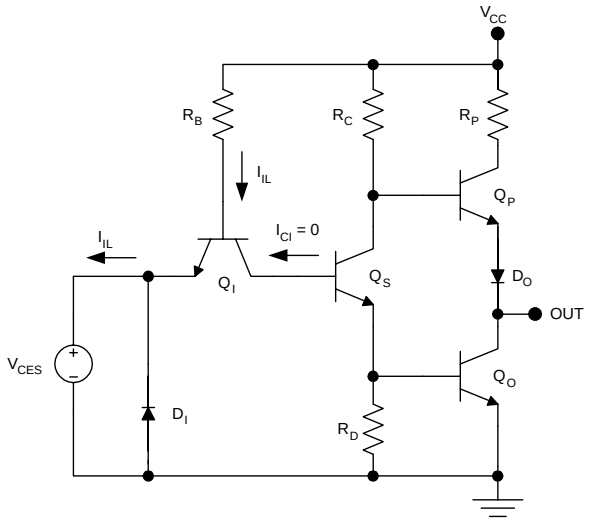


FIGURE 5.24
Standard TTL inverter for the calculation of I_{IL}.

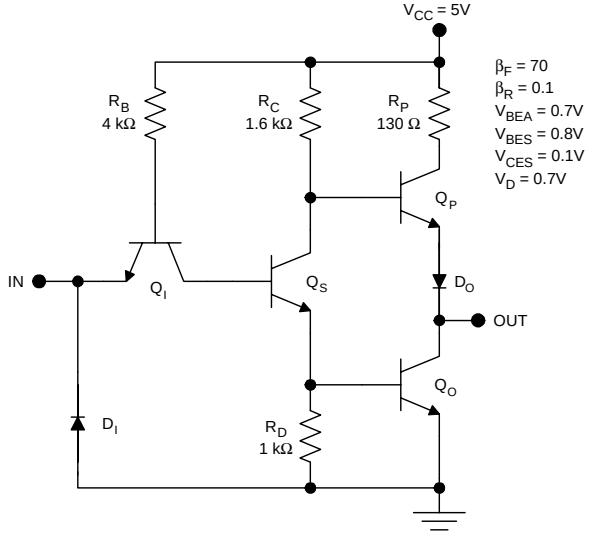


FIGURE 5.25
Standard TTL inverter for the example calculation of the DC fan-out.

Example 5.4

Calculate the maximum fan-out for the standard TTL inverter of Figure 5.25 based on DC considerations. Assume $\sigma_{MAX} = 1/2$.

Solution. The maximum current that the output can sink is

$$I_{OL} = \sigma_{MAX}\beta_F\left[(\beta_R + 1)\left(\frac{V_{CC} - V_{BCA} - 2V_{BES}}{R_B}\right) + \frac{V_{CC} - V_{CES} - V_{BES}}{R_C} - \frac{V_{BES}}{R_D}\right]$$

$$= (0.5)(70)\left[(1.1)\left(\frac{5\text{ V} - 0.7\text{ V} - 1.6\text{ V}}{4\text{ k}\Omega}\right) + \frac{5\text{ V} - 0.1\text{ V} - 0.8\text{ V}}{1.6\text{ k}\Omega} - \frac{0.8\text{ V}}{1\text{ k}\Omega}\right]$$

$$= 87.7\text{ mA}.$$

The input low current is

$$I_{IL} = \frac{V_{CC} - V_{BES} - V_{CES}}{R_B} = \frac{5\text{ V} - 0.8\text{ V} - 0.1\text{ V}}{4\text{ k}\Omega} = 1.025\text{ mA}.$$

The maximum fan-out based on DC considerations is the largest integer satisfying the inequality

$$N_{MAX} \le \frac{I_{OL}}{I_{IL}} = \frac{87.7\text{ mA}}{1.025\text{ mA}} = 85.6.$$

Thus,

$$N_{MAX} = 85.$$

5.7 Propagation Delays

Because the calculation of propagation delays for TTL circuits is rather complex, it is valuable first to consider the propagation delays for the simpler circuits from which TTL evolved.

First, the propagation delays for an unloaded bipolar transistor inverter will be considered. It will be demonstrated that the delays for this circuit stem from the RC time constants associated with the resistances in the circuit and the parasitic capacitances in the device. It will also be shown that the time required to bring a bipolar transistor out of saturation, the *saturation delay*, is the dominant speed limitation in the unloaded bipolar transistor inverter. Next, a bipolar transistor loaded by a lumped capacitive load will be considered. The delay associated with the product of the load capacitance and the collector resistor will be shown to be dominant in this case. This analysis justifies the use of active pull-up circuitry in TTL. Finally, the

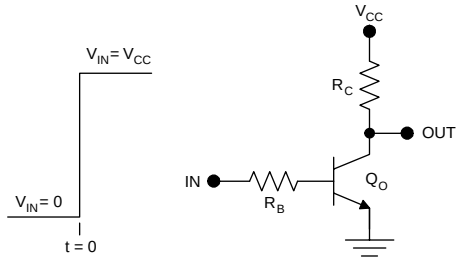


FIGURE 5.26
Unloaded RTL inverter with a low-to-high step drive at the input.

approximate analysis of the propagation delays in a simple TTL circuit will be considered. It will demonstrated that the saturation delay is still important in TTL, but that the propagation delays are relatively insensitive to capacitive loading.

5.7.1 Unloaded Transistor Inverter

Consider an unloaded RTL inverter with a step drive at the input as shown in Figure 5.26. If the input makes a low-to-high transition, then the output of the inverter will make a high-to-low transition. The associated propagation delay is t_{PHL}, the high-to-low propagation delay, and derives its name from the action at the *output* of the gate. The two contributions to t_{PHL} are the delay time, during which the transistor is cut off, and the fall time, during which the transistor is forward active.

During the delay time, the transistor is cut off and the base voltage increases from 0 to V_{BEA}. The delay time for the cutoff operation is

$$t_D = \frac{V_{BEA}\left(C_{BE} + C_{BC}\right)}{I_B(ave)}, \tag{5.33}$$

where C_{BE} and C_{BC} are the average junction capacitances and $I_B(ave)$ is the average base current. During the delay time, the base voltage increases from 0 to V_{BEA}. Therefore, the average base current during the delay time is

$$I_B(ave) = \frac{V_{CC} - V_{BEA}/2}{R_B} \text{ (delay time).} \tag{5.34}$$

The junction depletion capacitances C_{BE} and C_{BC} depend on the base-emitter and base-collector bias voltages, respectively. Therefore, both are also time dependent. For the purpose of hand calculations, the average junction depletion capacitance values can be used. The general procedure for any p–n junction is that, if the junction voltage varies from an initial value V_1 to a final value of V_2, the average junction capacitance is

$$C_J = \frac{C_{JO}V_{bi}}{(V_1 - V_2)(1-m)}\left[\left(1-\frac{V_2}{V_{bi}}\right)^{1-m} - \left(1-\frac{V_1}{V_{bi}}\right)^{1-m}\right], \tag{5.35}$$

where C_{JO} is the zero-bias capacitance, V_{bi} is the built-in potential for the junction, and m is the grading coefficient for the junction. The grading coefficient is a function of the doping gradient. For an abrupt junction, the grading coefficient is 1/2 and for a linearly graded junction the grading coefficient is 1/3. For typical junctions, $1/2 < m < 1/3$.

At the end of the delay time, the base-emitter voltage reaches V_{BEA} and the transistor becomes forward active. During the fall time, the transistor remains in the forward active mode; the output voltage falls as the collector current increases. The fall time may be calculated as

$$t_F = \frac{\Delta Q_F + \Delta V_{BC}C_{BC}}{I_B(ave)} = \frac{I_{CEOS}\tau_F + \Delta V_{BC}C_{BC}}{I_B(ave)}. \tag{5.36}$$

The first term in the numerator is the change in the minority carrier charge stored in the base of the transistor, where I_{CEOS} is the collector current at the edge of saturation and τ_F is the forward transit time for minority carriers in the base. The second term in the numerator is the change in the charge on the base-collector depletion capacitance, where ΔV_{BC} is the change in the base-collector voltage during the fall time and C_{BC} is the average base-collector depletion capacitance during the fall time. During the fall time, the base voltage is constant at V_{BEA}, so

$$I_B(ave) = \frac{V_{CC} - V_{BEA}}{R_B} \text{ (fall time).} \tag{5.37}$$

The high-to-low propagation delay is measured using the 50% points on the input and output waveforms. Therefore

$$t_{PHL} = t_D + \frac{t_F}{2}. \tag{5.38}$$

Now suppose the input voltage has been high for a long time and makes a high-to-low transition at $t = 0$ as shown in Figure 5.27. If the input makes a high-to-low transition, then the output of the inverter will make a low-to-high transition. The associated propagation delay is t_{PLH}, the low-to-high propagation delay, and it derives its name from the action at the *output* of the gate. The two contributions to t_{PLH} are the saturation delay, during which the transistor is saturated, and the rise time, during which the transistor is forward active.

Initially, the transistor is saturated. This means that both junctions in the transistor are forward biased, resulting in a lot of stored minority carrier

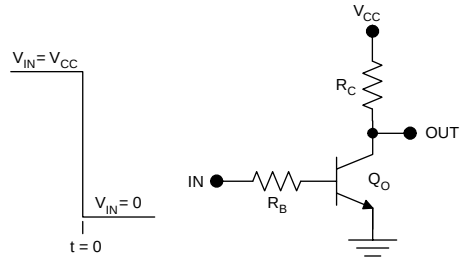


FIGURE 5.27
Unloaded RTL inverter with a high-to-low step drive at the input.

charge in the base and collector. The transistor remains saturated until most of this minority carrier charge is removed from the device. This delay is called the saturation delay and is given by

$$t_S = \tau_S \ln\left(\frac{I_{BF} - I_{BR}}{I_{CEOS}/\beta_F - I_{BR}}\right), \tag{5.39}$$

where τ_S is the saturation time constant, I_{BF} is the base current prior to $t = 0$, and I_{BR} is the base current during the saturation delay (after $t = 0$). Note that I_{BF} flows into the base and is positive, but I_{BR} flows out of the base and is negative. The saturation time constant may be calculated by

$$\tau_S = \frac{\alpha_F\left(\tau_F + \alpha_R\tau_R\right)}{1 - \alpha_F\alpha_R}, \tag{5.40}$$

where α_F and α_R are the forward and reverse common base current gain, respectively, and τ_F and τ_R are the forward and reverse effective lifetimes, respectively. At the end of the saturation delay, the base-collector voltage changes polarity and the transistor becomes forward active.

During the rise time, the transistor is forward active. The output voltage rises as the collector current decreases. The rise time may be calculated as

$$t_R = \frac{I_{CEOS}\tau_F + \left|\Delta V_{BC} C_{BC}\right|}{\left|I_B(ave)\right|}. \tag{5.41}$$

The absolute value function has been used to avoid negative values in the equation.

As with t_{PHL}, the low-to-high propagation delay is measured using the 50% points on the input and output waveforms. Therefore,

$$t_{PLH} = t_S + \frac{t_R}{2}. \tag{5.42}$$

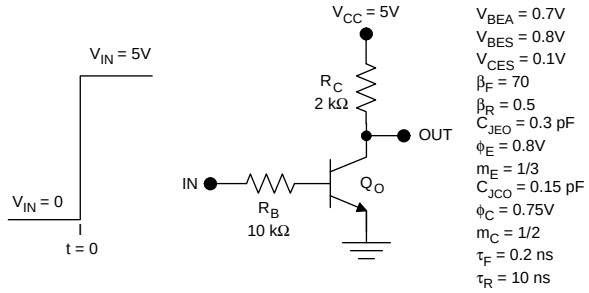


FIGURE 5.28
Unloaded RTL inverter for example calculation of t_{PHL}.

Example 5.5

Estimate the high-to-low propagation delay for the unloaded RTL inverter shown in Figure 5.28.

Solution. It is necessary to calculate the delay time and the fall time in order to obtain the high-to-low propagation delay. To determine the delay time, first calculate the average base current and the average values of C_{BE} and C_{BC}. The base voltage rises from 0 to 0.7 V, so the average base current is

$$I_B(ave) = \frac{V_{CC} - V_{BEA}/2}{R_B} = \frac{5\text{ V} - 0.7\text{ V}/2}{10\text{ k}\Omega} = 0.465\text{ mA}.$$

The base-emitter voltage increases from 0 to 0.7 V, so the average value of the base-emitter capacitance is

$$C_{BE} = \frac{C_{JEO}\phi_E}{(V_{BE1} - V_{BE2})(1 - m_E)}\left[\left(1 - \frac{V_{BE2}}{\phi_E}\right)^{1-m_E} - \left(1 - \frac{V_{BE1}}{\phi_E}\right)^{1-m_E}\right]$$

$$= \frac{(0.3\text{ pF})(0.8\text{ V})}{(0 - 0.7V)(0.667)}\left[\left(1 - \frac{0.7\text{ V}}{0.8\text{ V}}\right)^{0.667} - 1\right] = 0.386\text{ pF}.$$

The base-collector voltage starts at –5 V and ends up at –4.3 V at the end of the delay time. Therefore the average value of the base-collector capacitance for the delay time calculation is

$$C_{BC} = \frac{C_{JCO}\phi_C}{(V_{BC1} - V_{BC2})(1 - m_C)}\left[\left(1 - \frac{V_{BC2}}{\phi_C}\right)^{1-m_C} - \left(1 - \frac{V_{BC1}}{\phi_C}\right)^{1-m_C}\right]$$

$$= \frac{(0.15\text{ pF})(0.75\text{ V})}{(-5V - (-4.3\text{ V}))(0.5)}\left[\left(1 - \frac{-4.3\text{ V}}{0.75\text{ V}}\right)^{0.5} - \left(1 - \frac{-5\text{ V}}{0.75\text{ V}}\right)^{0.5}\right] = 0.0559\text{ pF}.$$

The delay time is therefore

$$t_D = \frac{V_{BEA}\left(C_{BE} + C_{BC}\right)}{I_B(ave)} = \frac{0.7\text{ V}\left(0.386\text{ pF} + 0.0559\text{ pF}\right)}{0.465\text{ mA}} = 0.665\text{ ns}.$$

For the calculation of the fall time, the base-collector voltage starts at –4.3 V and ends up at +0.7 V. Therefore, the average value of the base-collector capacitance for the fall time calculation is

$$C_{BC} = \frac{C_{JCO}\phi_C}{\left(V_{BC1} - V_{BC2}\right)\left(1 - m_C\right)}\left[\left(1 - \frac{V_{BC2}}{\phi_C}\right)^{1-m_C} - \left(1 - \frac{V_{BC1}}{\phi_C}\right)^{1-m_C}\right]$$

$$= \frac{\left(0.15\text{ pF}\right)\left(0.75\text{ V}\right)}{\left(-4.3V - \left(0.7\text{ V}\right)\right)\left(0.5\right)}\left[\left(1 - \frac{0.7\text{ V}}{0.75\text{ V}}\right)^{0.5} - \left(1 - \frac{-4.3\text{ V}}{0.75\text{ V}}\right)^{0.5}\right] = 0.1052\text{ pF}.$$

Also,

$$I_{CEOS} = \frac{V_{CC} - V_{CES}}{R_C} = \frac{5\text{ V} - 0.1\text{ V}}{2\text{ k}\Omega} = 2.45\text{ mA}$$

and, for the fall time calculation,

$$I_B(ave) = \frac{V_{CC} - V_{BEA}}{R_B} = \frac{5\text{ V} - 0.7\text{ V}}{10\text{ k}\Omega} = 0.43\text{ mA}.$$

Therefore, the fall time is

$$t_F = \frac{I_{CEOS}\tau_F + \Delta V_{BC} C_{BC}}{I_B(ave)}$$

$$= \frac{\left(2.45\text{ mA}\right)\left(0.2\text{ ns}\right) + \left(0.7\text{ V} - \left(-4.3\text{ V}\right)\right)\left(0.1052\text{ pF}\right)}{0.43\text{ mA}} = 2.36\text{ ns}.$$

The unloaded high-to-low propagation delay is therefore

$$t_{PHL} = t_D + \frac{t_F}{2} = 0.665\text{ ns} + \frac{2.36\text{ ns}}{2} = 1.845\text{ ns}.$$

Example 5.6

Estimate the low-to-high propagation delay for the unloaded RTL inverter shown in Figure 5.29.

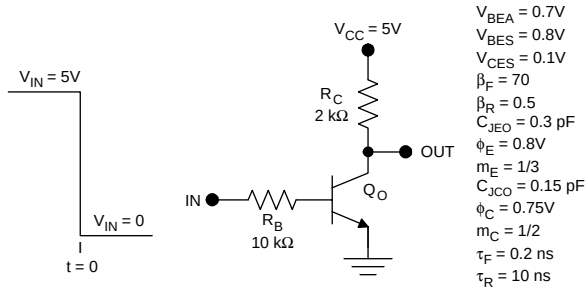


FIGURE 5.29
Unloaded RTL inverter for the calculation of t_{PLH}.

Solution. The saturation time constant is

$$\tau_S = \frac{\alpha_F\left(\tau_F + \alpha_R\tau_R\right)}{1-\alpha_F\alpha_R} = \frac{0.986\left(0.2\text{ ns} + (0.333)10\text{ ns}\right)}{1-(0.986)(0.333)} = 5.2\text{ ns}.$$

The saturated base current is

$$I_{BF} = \frac{V_{CC} - V_{BES}}{R_B} = \frac{5\text{ V} - 0.8\text{ V}}{10\text{ k}\Omega} = 0.42\text{ mA}.$$

During the saturation delay, the base current is

$$I_{BR} = \frac{0 - V_{BES}}{R_B} = \frac{0 - 0.8\text{ V}}{10\text{ k}\Omega} = -0.08\text{ mA}.$$

The collector current at the edge of saturation is

$$I_{CEOS} = \frac{V_{CC} - V_{CES}}{R_C} = \frac{5\text{ V} - 0.1\text{ V}}{2\text{ k}\Omega} = 2.45\text{ mA}.$$

The saturation delay is therefore

$$t_S = \tau_S \ln\left(\frac{I_{BF} - I_{BR}}{I_{CEOS}/\beta_F - I_{BR}}\right) = 5.2\text{ ns}\ln\left(\frac{0.42\text{ mA} - (-0.08\text{ mA})}{2.45\text{ mA}/70 - (-0.08\text{ mA})}\right)$$

$$= 7.6\text{ ns}.$$

For calculation of the rise time, the base-collector voltage starts at +0.7 V and ends up at –4.3 V. Therefore the average value of the base-collector capacitance for the rise time calculation is

$$C_{BC} = \frac{C_{JCO}\phi_C}{(V_{BC1} - V_{BC2})(1 - m_C)}\left[\left(1 - \frac{V_{BC2}}{\phi_C}\right)^{1-m_C} - \left(1 - \frac{V_{BC1}}{\phi_C}\right)^{1-m_C}\right]$$

$$= \frac{(0.15\text{ pF})(0.75\text{ V})}{(0.7V - (-4.3V))(0.5)}\left[\left(1 - \frac{-4.3\text{ V}}{0.75\text{ V}}\right)^{0.5} - \left(1 - \frac{0.7\text{ V}}{0.75\text{ V}}\right)^{0.5}\right] = 0.1052\text{ pF}.$$

The base current during the rise time is

$$I_B(ave) = \frac{0 - V_{BEA}}{R_B} = \frac{0 - 0.7\text{ V}}{10\text{ k}\Omega} = -0.07\text{ mA}.$$

The rise time is

$$t_R = \frac{I_{CEOS}\tau_F + |\Delta V_{BC} C_{BC}|}{|I_B(ave)|} = \frac{(2.45\text{ mA})0.2\text{ ns} + |(-4.3\text{ V} - 0.7\text{ V})(0.1052\text{ pF})|}{|-0.07\text{ mA}|}$$

$$= 14.51\text{ ns}$$

and the low-to-high propagation delay is

$$t_{PLH} = t_S + \frac{t_R}{2} = 7.6\text{ ns} + \frac{14.51\text{ ns}}{2} = 15.3\text{ ns}.$$

Two important points should be noted here. First, the low-to-high propagation delay is many times longer than the high-to-low propagation delay. Second, t_{PLH} is often dominated by the saturation delay. These results are true in general for RTL circuitry and show that any effort toward speed improvement should be directed toward the prevention of saturation in the switch transistors.

5.7.2 Loaded Transistor Inverter

Here the load is assumed to be a pure capacitance placed between the output of the gate and ground as shown in Figure 5.30. In addition, this load capacitance is assumed to be much larger than the parasitic capacitances in the transistors. This approximation is relevant for the case of an off-chip load. The case of a lumped capacitive load is sufficiently complicated that accurate performance predictions require computer simulations. However, approximate hand calculations can be made using reasonable approximations.

The high-to-low propagation delay is approximately unchanged from the unloaded case because the delays inherent in the transistor dominate. Of these, the fall time associated with C_{BC} is important because of the Miller effect associated with this feedback capacitor. On the other hand, the external load capacitance can be discharged rapidly due to the high current gain of

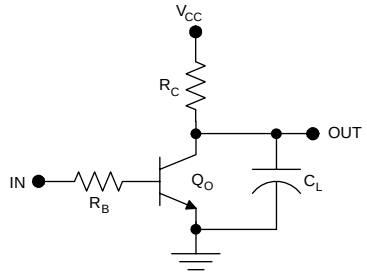


FIGURE 5.30
RTL inverter with a lumped capacitive load.

the transistor. The low-to-high propagation delay is lengthened considerably by the addition of an external capacitive load; this delay can be estimated as the saturation delay plus one half the rise time. The saturation delay is calculated exactly as before and is given by

$$t_S = \tau_S \ln\left(\frac{I_{BF} - I_{BR}}{I_{CEOS}/\beta_F - I_{BR}}\right). \tag{5.43}$$

Determination of the rise time is very different from determining the unloaded case, however. The assumption here is that it takes a negligible time for the transistor to cut off after the saturation delay. Then only the rise time for a simple RC circuit comprising C_L and R_C needs to be calculated:

$$V_{\text{OUT}}(t) = V_{CES} + \left(V_{CC} - V_{CES}\right)\left(1 - \exp\left(-\frac{t - t_S}{R_C C_L}\right)\right). \tag{5.44}$$

Using the 50% point in the output waveform, the low-to-high propagation delay can be determined:

$$t_{PLH} = t_S + \ln(2) R_C C_L \text{ (lumped capacitive load).} \tag{5.45}$$

Example 5.7

Estimate the low-to-high propagation delay for the RTL inverter with a lumped capacitive load as depicted in Figure 5.31.

Solution. Using the values of I_{BF}, I_{BR}, and I_{CEOS} calculated in Example 5.6, the saturation delay is

$$t_S = \tau_S \ln\left(\frac{I_{BF} - I_{BR}}{I_{CEOS}/\beta_F - I_{BR}}\right) = 5.2 \text{ ns} \ln\left(\frac{0.42 \text{ mA} - (-0.08 \text{ mA})}{2.45 \text{ mA}/70 - (-0.08 \text{ mA})}\right)$$

$$= 7.6 \text{ ns}.$$

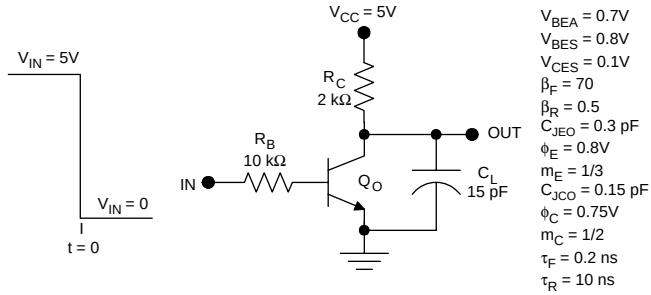


FIGURE 5.31
Example RTL inverter with a lumped capacitive load for the calculation of t_{PLH}.

The low-to-high approximate propagation delay is

$$t_{PLH} \approx t_S + \ln(2)R_C C_L = 7.6 \text{ ns} + (0.693)(2 \text{ k}\Omega)(15 \text{ pF}) = 28 \text{ ns}.$$

Notice that t_{PLH} is much greater than t_{PHL}, as with the unloaded case. Notice too that the rise time is greater than the saturation delay. This is very different from the unloaded situation.

5.7.3 Loaded TTL Inverter

The addition of active pull-up circuitry greatly improves the dynamic performance of TTL over the other saturated bipolar logic circuits. However, it is still true that $t_{PLH} \gg t_{PHL}$ because of the output transistor saturation delay.

For the estimation of t_{PLH}, suppose a high-to-low step drive is applied at the input of the standard TTL inverter as shown in Figure 5.32. Once again several contributions are made to the overall delay. First, the input transistor must become forward active. This occurs very rapidly and the delay can be neglected for the purpose of hand calculations. Second, Q_S must switch off. The saturation delay associated with Q_S is very short and can be neglected because the forward active Q_I will rapidly remove base charge from Q_S. Third, the output transistor Q_O must come out of saturation and switch off. Fourth, the pull-up transistor must bring the output voltage from low to high. The first two delays will be neglected; therefore, the low-to-high propagation delay is equal to the saturation delay for Q_O plus one half the rise time for the output voltage:

$$t_{PLH} \approx t_S + \frac{t_R}{2}. \tag{5.46}$$

The saturation delay for the output transistor can be calculated in the usual manner and is given by

$$t_S = \tau_S \ln\left(\frac{I_{BFO} - I_{BRO}}{I_{CEOS}/\beta_F - I_{BRO}}\right). \tag{5.47}$$

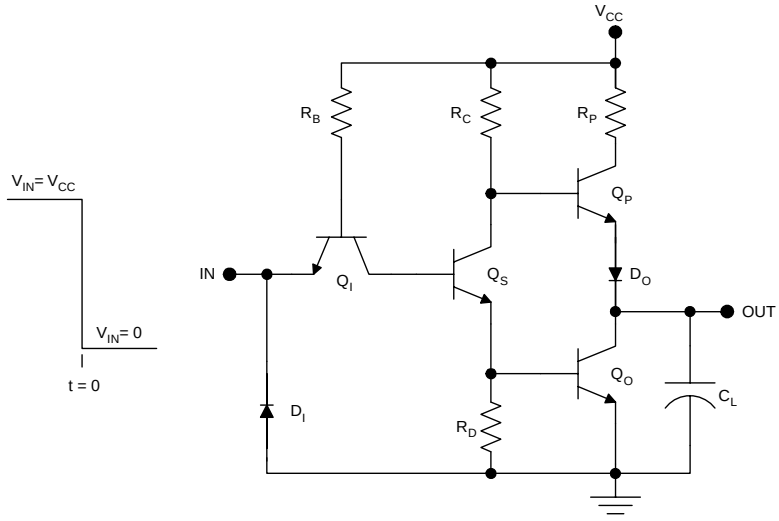


FIGURE 5.32
Standard TTL inverter for the estimation of t_{PLH}.

The rise time at the output is very short due to the low output impedance of the pull-up circuit. For this emitter follower, the output impedance is

$$R_{\text{OUT}} = \frac{R_C}{\beta_F} + \frac{kT}{qI_{EP}}, \tag{5.48}$$

where kT/q is the thermal voltage (26 mV at room temperature) and I_{EP} is the emitter current in the pull-up transistor. As the output current varies with time, so too will the output impedance. For the purpose of an approximate hand calculation, use the initial value of I_{EP}. This initial value of I_{EP} flows when the output voltage is still at V_{OL} and is given by

$$I_{EP} = (\beta_F + 1)\left(\frac{V_{CC} - V_{BEA} - V_D - V_{CES}}{R_C}\right) \approx \beta_F\left(\frac{V_{CC} - V_{BEA} - V_D - V_{CES}}{R_C}\right). \tag{5.49}$$

Therefore, the initial output impedance of the pull-up circuit is

$$R_{\text{OUT}} = \frac{R_C}{\beta_F}\left[1 + \frac{V_T}{V_{CC} - V_{BEA} - V_D - V_{CES}}\right] \approx \frac{R_C}{\beta_F}. \tag{5.50}$$

This approximation results from the fact that the right-hand quantity in the brackets is much smaller than one. The approximate rise time can be calculated by analogy to a simple RC circuit and is

$$t_R \approx \ln 2 \frac{R_C C_L}{\beta_F} . \tag{5.51}$$

The rise time will be very short in comparison to the case of the passive pull-up in the RTL circuit because β_F appears in the denominator while all other quantities are similar to the RTL case. Therefore, the rise time is reduced by a factor of 50 or more compared to the passive pull-up case, and the saturation delay for Q_O becomes the dominant contribution to t_{PLH}. Typically, the rise time will be less than 1 ns, whereas the saturation delay will be 10 ns. Therefore,

$$t_{PLH} \approx \tau_S \ln\left(\frac{I_{BFO} - I_{BRO}}{I_{CEOS}/\beta_F - I_{BRO}} \right), \tag{5.52}$$

where

$$I_{BFO} = \left(1 + \beta_R\right) \frac{V_{CC} - V_{BCA} - 2V_{BES}}{R_B} + \frac{V_{CC} - V_{CES} - V_{BES}}{R_C} - \frac{V_{BES}}{R_D} , \tag{5.53}$$

$$I_{BRO} = -\frac{V_{BES}}{R_D} , \tag{5.54}$$

and

$$I_{CEOS} = NI_{IL} . \tag{5.55}$$

Example 5.8

Estimate the low-to-high propagation delay for the standard TTL inverter of Figure 5.33. Assume $N = 10$, and $\tau_S = 20$ ns.

Solution. For $t < 0$, the forward base current for Q_O is

$$I_{BF} = \left(\beta_R + 1\right)\left(\frac{V_{CC} - V_{BCA} - 2V_{BES}}{R_B} \right) + \frac{V_{CC} - V_{CES} - V_{BES}}{R_C} - \frac{V_{BES}}{R_D}$$

$$= (1.1)\left(\frac{5 \text{ V} - 0.7 \text{ V} - 1.6 \text{ V}}{4 \text{ k}\Omega} \right) + \frac{5 \text{ V} - 0.1 \text{ V} - 0.8 \text{ V}}{1.6 \text{ k}\Omega} - \frac{0.8 \text{ V}}{1 \text{ k}\Omega} = 2.5 \text{ mA};$$

during the saturation delay, the Q_O base current is

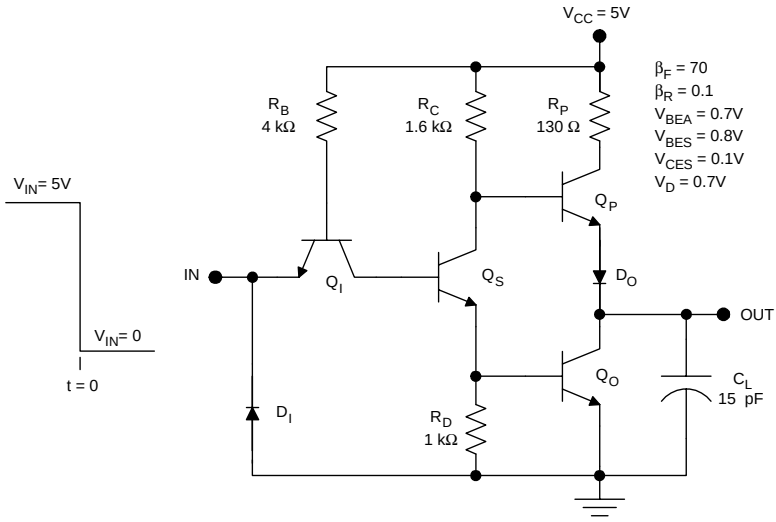


FIGURE 5.33
Standard TTL inverter for example calculation of t_{PLH}.

$$I_{BR} = -\frac{V_{BES}}{R_D} = -\frac{0.8\text{ V}}{1\text{ k}\Omega} = -0.08\text{ mA}$$

and the collector current at the edge of saturation is

$$I_{CEOS} = NI_{IL} = N\left(\frac{V_{CC} - V_{BES} - V_{CES}}{R_B}\right) = (10)\left(\frac{5\text{ V} - 0.8\text{ V} - 0.1\text{ V}}{4\text{ k}\Omega}\right)$$

$$= 10.25\text{ mA.}$$

The saturation delay is

$$t_S = \tau_S \ln\left(\frac{I_{BF} - I_{BR}}{I_{CEOS}/\beta_F - I_{BR}}\right) = 20\text{ ns}\ln\left(\frac{2.5\text{ mA} - (-0.08\text{ mA})}{10.25\text{ mA}/70 - (-0.08\text{ mA})}\right)$$

$$= 49\text{ ns}$$

and the low-to-high propagation delay is approximately the same

$$t_{PLH} \approx t_S = 49\text{ ns}.$$

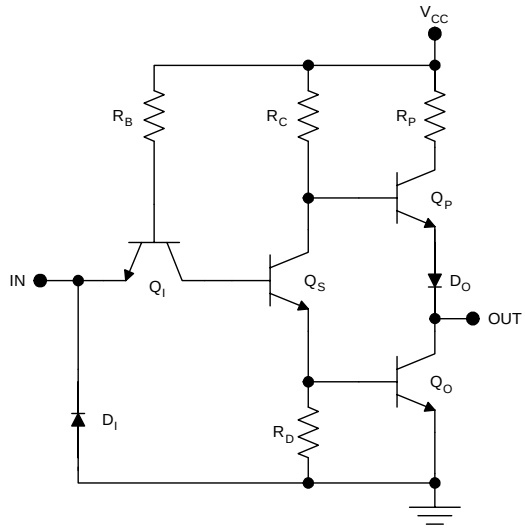


FIGURE 5.34
TTL inverter.

5.8 Logic Design

The standard TTL gates discussed so far are inverting. Thus, the one-input gate shown in Figure 5.34 is an inverter. The NAND gate is constructed by using an input transistor with multiple emitters as shown in Figure 5.35. This multi-emitter input transistor performs the ANDing of the input signals because it will saturate if any one of the inputs goes low. Only in the case of all high inputs will the input transistor become reverse active; in other words, forward bias of any one base–emitter junction will result in the injection of minority carriers into the base. As a result, the input transistor performs the AND function. In addition, the drive splitter is a noninverting emitter follower and the output transistor is an inverting common emitter amplifier, so the overall function is NAND. Although the circuit shown is a three-input version, it is entirely possible to fabricate such a NAND gate with up to eight inputs.

The NOR gate is realized by the use of multiple input transistors and drive splitters as shown in Figure 5.36. The operation of the circuit is as follows. If V_A goes high, then Q_{IA} becomes reverse active and Q_{SA} saturates. Then, regardless of the conduction state of Q_{SB}, Q_P will be cut off and Q_O will saturate, bringing the output low. The same reasoning applies to V_B. Therefore, if either input goes high, the output goes low, resulting in the NOR function.

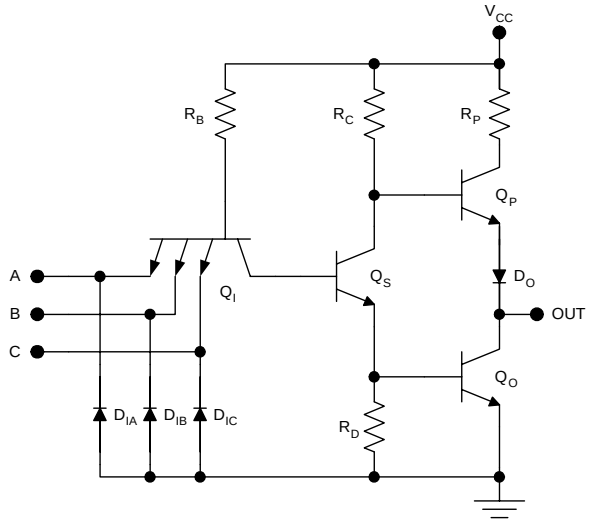


FIGURE 5.35
TTL NAND3 gate.

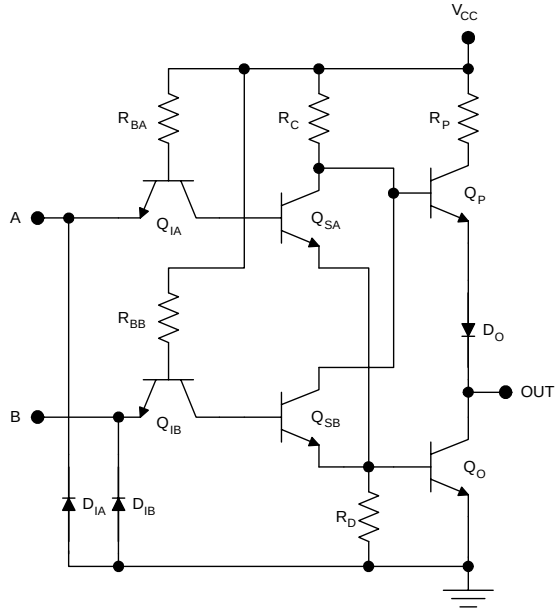


FIGURE 5.36
TTL NOR2 gate.

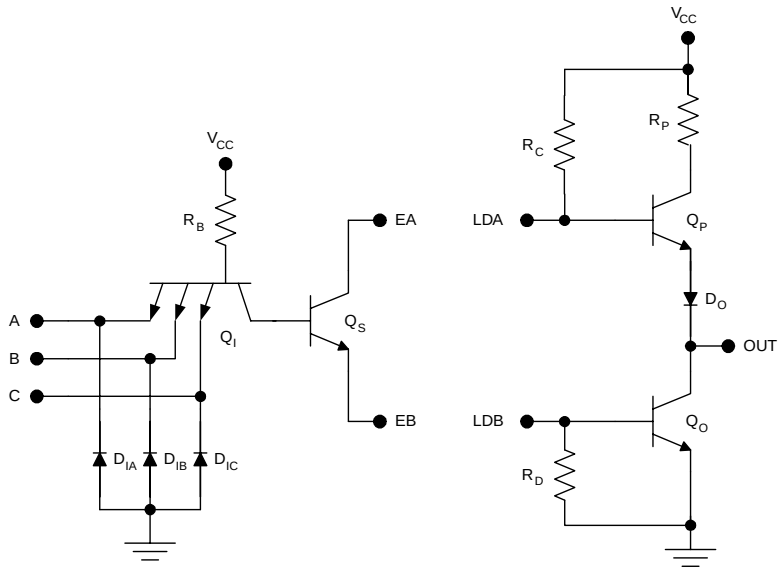


FIGURE 5.37
TTL expander (left) and line driver (right).

General AND–OR–INVERT functions can be created in TTL by the connection of expanders and line drivers. These TTL subcircuits, shown in Figure 5.37, are available separately and are user connected to achieve the desired logic function. The expander includes the input transistor and the drive splitter. Drive splitters are available with various numbers of inputs ranging from one to eight. The line driver comprises only the totem pole output and allows the use of active pull-up circuitry in user-configured logic gates.

An AND–OR–INVERT gate is user configured by the connection of expanders and line drivers as shown in Figure 5.38. The example shown is called a "three input, two wide" AND–OR–INVERT gate; the input transistors form the AND function. ORing is achieved by parallel connection of drive splitters, and the totem pole output is inherently inverting. Therefore, the Boolean function realized at the output of the circuit shown is

$$Y = \overline{(A+B+C)\bullet(D+E+F)}\,. \tag{5.56}$$

The exclusive OR (XOR) function is of special interest because of its importance in adders. Figure 5.39 shows the implementation of the XOR function in standard TTL. The key subcircuit is the pair of transistors Q_{X1} and Q_{X2} (the "exclusive OR pair"). Suppose A is high and B is low. Under these conditions, Q_{IA} is reverse active and Q_{SA} and Q_{SDA} are saturated. This brings the emitter of Q_{X2} to V_{CES}, or about 1/10 of a volt. At the same time, Q_{IB} is saturated but Q_{SB} and Q_{SDB} are cut off; this brings the base of Q_{X2} high. With

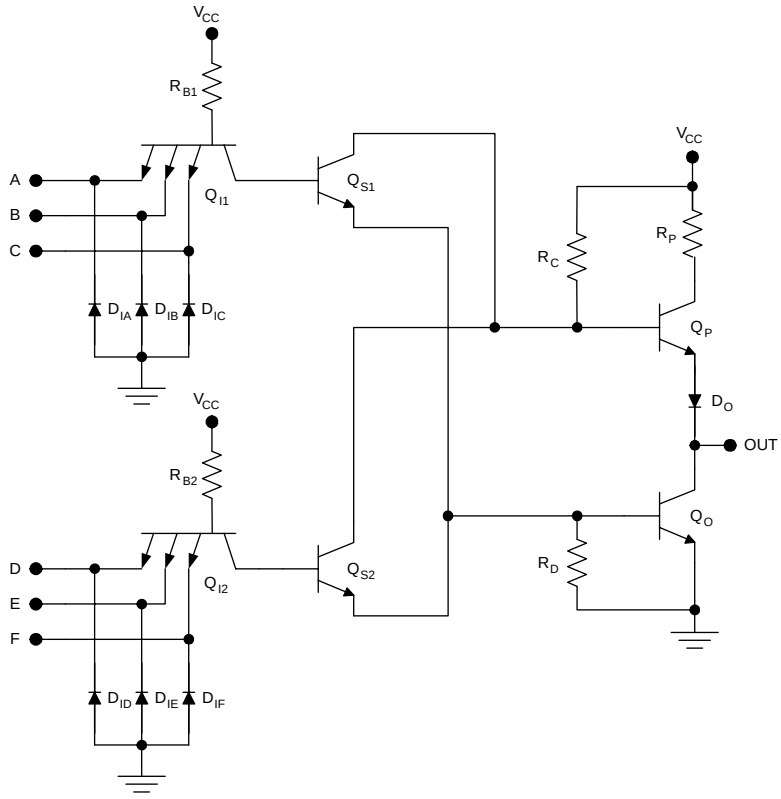


FIGURE 5.38
TTL three-input, two-wide AND–OR–INVERT gate.

A high and B low, Q_{X2} saturates, causing Q_S and Q_O to cut off; as a result, the output goes high. Similarly, if A is low and B is high, Q_{X1} saturates and the output goes high. However, if both inputs are low or both are high, both transistors in the XOR pair experience zero base-emitter voltage difference and cut off. Under these conditions, Q_S and Q_O saturate, bringing the output low. Therefore, the resulting logic function is XOR.

The logic design concepts discussed here have all been illustrated using standard 74xx series TTL circuitry for simplicity. However, the same logic design concepts can be applied to the more modern TTL families with subtle changes.

5.9 Schottky TTL

Schottky TTL gate circuits outperform standard TTL due to a number of device and circuit improvements. Of all the circuit improvements, three

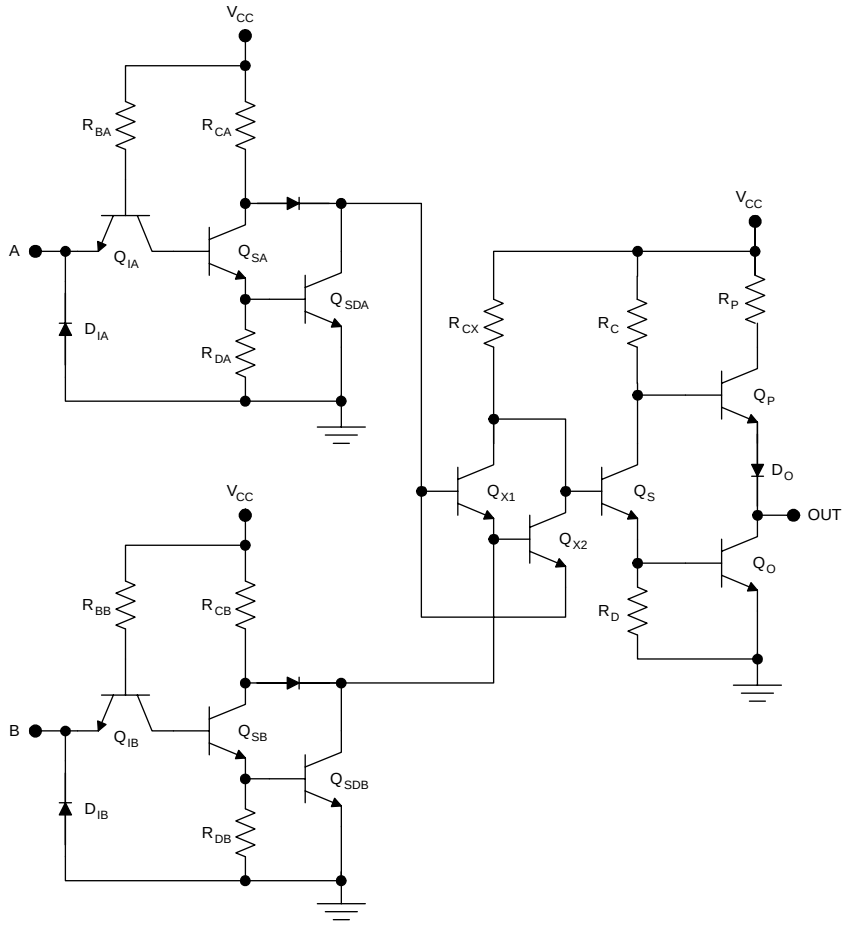


FIGURE 5.39
TTL XOR2 gate.

stand out above the rest: the use of Schottky clamping, pseudo Darlington pull-up subcircuits, and squaring subcircuits.

5.9.1 Schottky Clamping

In 74xx series TTL, the low-to-high propagation delay is dominated by the saturation delay for the output transistor. Generally speaking, the elimination of saturation delays for all of the transistors can greatly improve t_{PLH} and, to a lesser extent, t_{PHL}.

A straightforward way of eliminating saturation is to use Schottky clamping, in which a metal semiconductor (Schottky) diode is placed between the base and collector of each bipolar transistor as shown in Figure 5.40. The Schottky diode is inserted with its anode at the transistor base and thus only

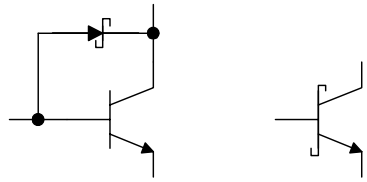


FIGURE 5.40
Schottky clamped transistor (left) and circuit symbol (right).

conducts when the base–collector junction is forward biased. Although the Schottky diode follows the normal diode equation, it has a much smaller turn-on voltage than the base–collector junction of the bipolar transistor. Under forward active conditions, the base–collector junction and the Schottky diode are reverse biased. Therefore the diode has no effect on the circuit operation and the full current gain of the transistor is available.

Hard saturation of the Schottky clamped transistor is prevented because the diode shunts excess base drive current to the collector before the base–collector junction is sufficiently biased to cause significant injection of excess minority carriers into the base of the transistor. As the base drive is increased, V_{CE} decreases in order to maintain the forward active current relationship $I_C = \beta_F I_B$. However, the clamp diode prevents V_{CE} from reaching the value V_{CES}. Instead, the minimum value of the collector emitter voltage is reached in a new "mode" of operation called "on hard." This minimum value of V_{CE} is

$$V_{CEOH} = V_{BEOH} - V_{SBD}\,, \tag{5.57}$$

where V_{BEOH} is the base emitter voltage corresponding to on-hard operation and V_{SBD} is the turn-on voltage for the *Schottky barrier diode*. Typically, V_{BEOH} is similar to the saturated value of the base emitter voltage or about 8/10 of a volt. The turn-on voltage for the Schottky diode is typically 3/10 of a volt for a silicon–platinum silicide Schottky diode. Therefore, V_{CEOH} is typically 1/2 volt.

Strictly speaking, the transistor is saturated under on-hard conditions because both junctions are forward biased; however, there are two reasons for calling this operation "on hard" rather than "saturation." First, although the base–collector junction is forward biased, the bias voltage is only about 3/10 of a volt. This results in negligible minority carrier injection at the base–collector junction, so negligible saturation charge is in the base or collector. Second, because of negligible injection from the collector side of the base, the forward active current relationship is maintained. Thus, the on-hard mode of operation is unlike hard saturation in terms of the stored charge and the current relationships. On the other hand, it is unlike forward active operation in that the collector emitter voltage is clamped at a minimum value. This unique behavior justifies the use of a new name for this "mode" of operation.

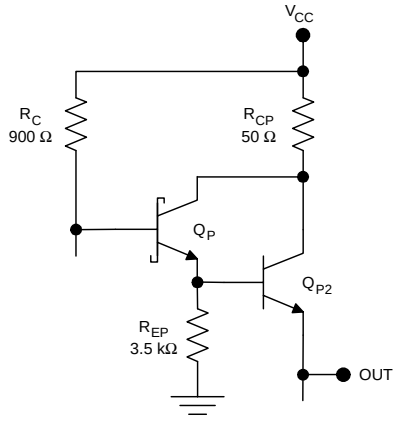


FIGURE 5.41
Pseudo Darlington pull-up subcircuit.

5.9.2 Pseudo Darlington Pull-Up Subcircuit

The pseudo Darlington pull-up subcircuit (Figure 5.41) is another refinement included universally in modern versions of TTL. The essential design idea is to reduce the output impedance of the pull-up circuit, thus reducing the low-to-high propagation delay for the case of a highly capacitive load. Here, the pseudo Darlington pair provides very low output impedance (1 Ω) in the emitter follower configuration. It is not a true Darlington pair because of the inclusion of R_{EP}. Therefore, the full emitter current of Q_P does not flow to the base of Q_{P2}; however, R_{EP} is necessary to drain base charge from Q_{P2}, thus allowing this device to turn off rapidly during high-to-low transitions at the output.

The pseudo Darlington pull-up circuit drops the series diode used in the totem pole output because it was necessary only to ensure that the pull-up and pull-down transistors did not conduct simultaneously under DC conditions. However, the voltage drop across the base–emitter junction of the second transistor obviates the need for this diode in the pseudo Darlington configuration. Note also that the removal of this diode avoids any degradation in V_{OH} with the addition of a second transistor.

It is unnecessary to clamp Q_{P2} with a Schottky diode because Q_{P2} is inherently clamped by Q_P and cannot saturate. The collector–base junction of Q_{P2} experiences a minimum reverse bias equal to V_{CEOH} for Q_P. The pseudo Darlington subcircuit shown here is the version used in 74S/54S Schottky TTL. However, the pull-up subcircuits used in other TTL families are similar in concept.

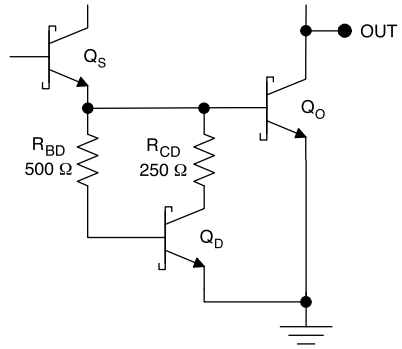


FIGURE 5.42
Squaring subcircuit.

5.9.3 Squaring Subcircuit

All modern versions of TTL use a squaring subcircuit as shown in Figure 5.42. This subcircuit influences the DC and the dynamic performance of the gate circuit as discussed next.

The squaring subcircuit eliminates the first breakpoint of the voltage transfer characteristic for standard TTL because there is no path for the conduction of emitter current from the drive splitter Q_S until the base-emitter voltage of Q_O reaches V_{BEA}. At this point, Q_S, Q_O, and Q_D turn on together. The elimination of the first breakpoint in the VTC raises V_{IL} and improves the low noise margin. It also "squares" the voltage transfer characteristic (makes it less rounded in appearance).

The squaring subcircuit also improves the dynamic performance of the circuit. During low-to-high transitions at the output, the squaring circuit allows faster removal of charge from the base of Q_O than a simple pull-down resistor. This allows Q_O to turn off faster, improving t_{PLH} and reducing the dynamic power dissipation. Because of this, the subcircuit is sometimes said to provide "active pull-down" for the base of the output transistor. The version shown in Figure 5.42 is from the 74S/54S Schottky TTL family, but other versions are conceptually the same.

5.9.4 74S Schottky TTL (STTL)

The simplest version of a Schottky clamped TTL is the 74S/54S series of parts, simply known as "Schottky TTL." Figure 5.43 shows the three-input NAND gate. Using the 74/54 standard TTL as a point of comparison, five key refinements improve the power delay product: 1) the use of Schottky clamping for all transistors except Q_{P2}; 2) the incorporation of pseudo Darlington pull-up; 3) the inclusion of a squaring subcircuit; 4) improved

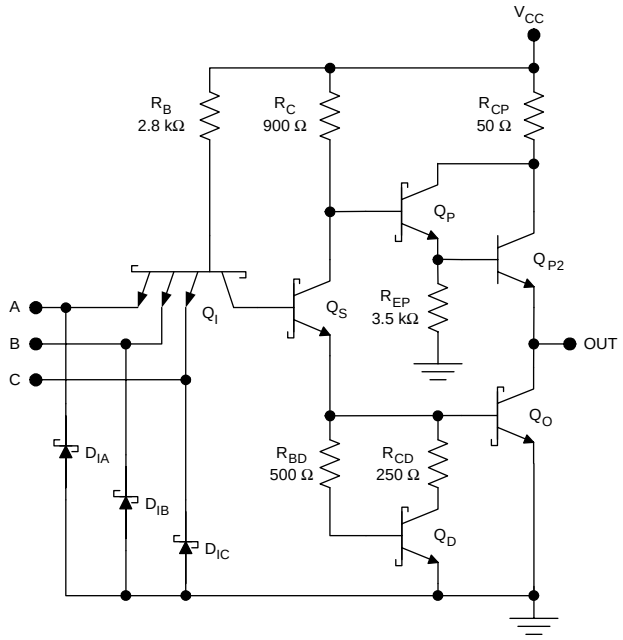


FIGURE 5.43
STTL NAND3 gate.

transistor fabrication resulting in smaller devices with reduced parasitic capacitances; and 5) the use of Schottky antiringing diodes. The absence of minority carrier storage effects in the Schottky antiringing diodes makes them faster than their p–n junction counterparts. One other circuit modification is the use of scaled down resistors. This reduces RC products in the circuit and improves the speed of the circuit, but at the expense of increased dissipation. Thus the scaling of the resistors involves a direct trade-off between speed and power rather than an improvement in the power delay product.

Example 5.9

Determine the voltage transfer characteristic for the Schottky TTL inverter of Figure 5.44.

Solution. With the input low, Q_I is on hard, Q_S, Q_D, and Q_O are cut off, and the output goes high. Both pull-up transistors are forward active so that

$$V_{OH} = V_{CC} - 2V_{BEA} = 5\text{ V} - 1.4\text{ V} = 3.6\text{ V}\,.$$

If the input voltage is increased, Q_S, Q_O, and Q_D will turn on (become forward active) simultaneously. The corresponding input voltage is

$$V_{IL} = 2V_{BEA} - V_{CEOH} = 1.4\text{ V} - 0.5\text{ V} = 0.9\text{ V}\,;$$

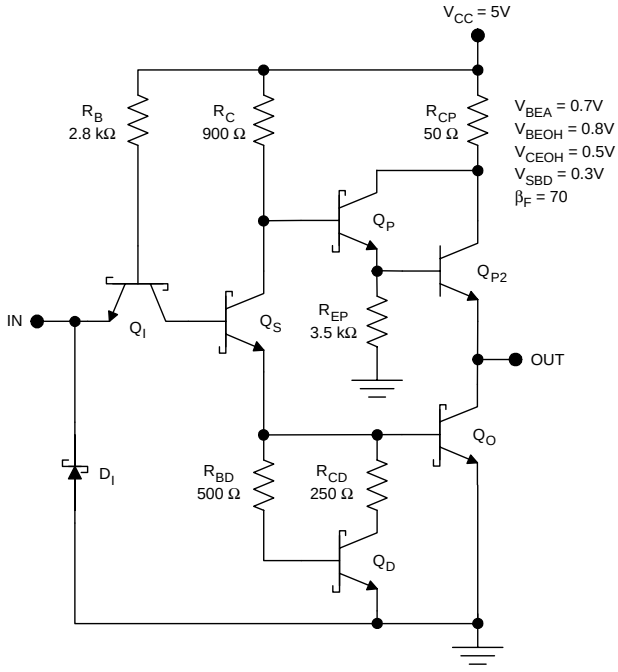


FIGURE 5.44
STTL inverter.

Q_S, Q_D, and Q_O will be on hard if input voltage is increased further to

$$V_{IH} = 2V_{BEOH} - V_{CEOH} = 1.6\text{ V} - 0.5\text{ V} = 1.1\text{ V}.$$

The corresponding output voltage is

$$V_{OL} = V_{CEOH} = 0.5\text{ V}.$$

Figure 5.45 shows the voltage transfer characteristics for 74S/54S Schottky TTL and 74/54 standard TTL for the sake of comparison. Although standard TTL exhibits three breakpoints, Schottky TTL has a squared characteristic with only two breakpoints, as expected.

Example 5.10

Calculate I_{IL} for the Schottky TTL gate of Figure 5.46.

Solution. With logic zero input from a similar gate, Q_I is on hard and Q_S is cut off. Therefore,

$$I_{IL} = \frac{V_{CC} - V_{BEOH} - V_{OL}}{R_B} = \frac{5\text{ V} - 0.8\text{ V} - 0.5\text{ V}}{2.8\text{ k}\Omega} = 1.32\text{ mA}.$$

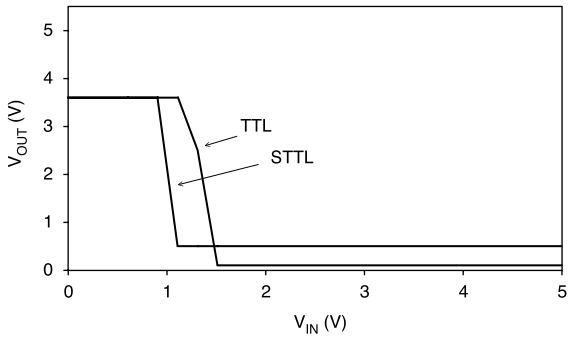


FIGURE 5.45
Voltage transfer characteristics for 74/54 TTL and 74S/54S Schottky TTL.

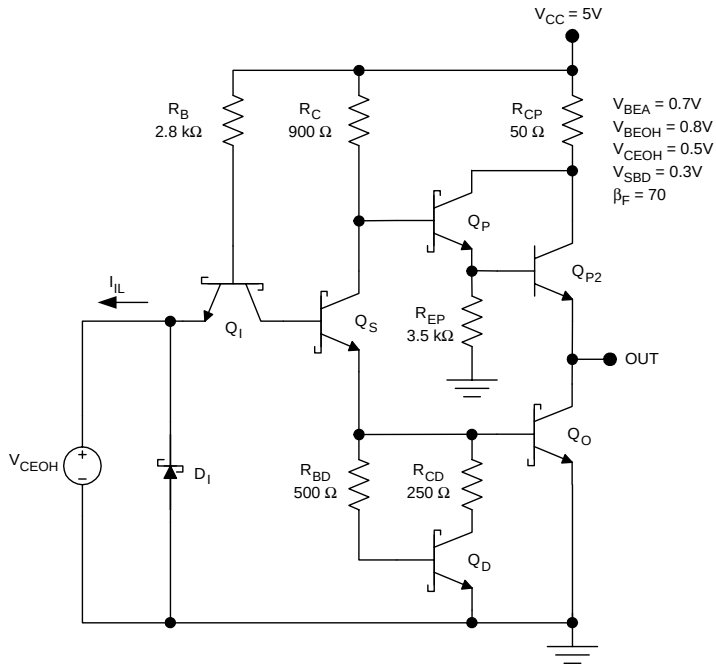


FIGURE 5.46
STTL inverter for the calculation of I_{IL}.

Example 5.11

Determine I_{IH} for the Schottky TTL gate of Figure 5.47.

Solution. If a logic-one level is applied from a similar gate, then the base–emitter junction of Q_I is reverse biased but the base–collector junction is forward biased. If the definitions are strictly applied, this is reverse-active operation; however, the normal reverse-active current relationship does not hold because the Schottky clamp diode conducts, shunting current from the

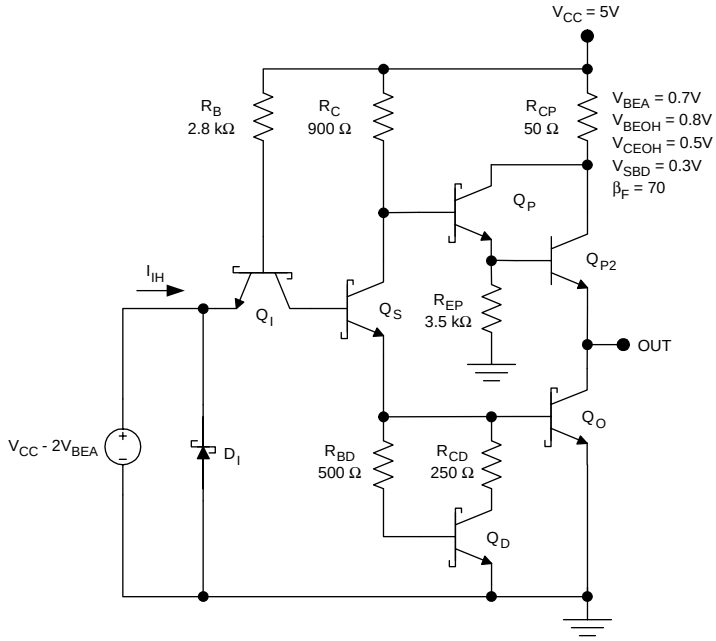


FIGURE 5.47
STTL inverter for the calculation of I_{IH}.

base to the collector of the transistor. This new mode of operation is sometimes called "reverse Schottky."

With the input transistor operating in the reverse Schottky mode, negligible minority carrier charge is injected into the base by the collector and therefore no significant reverse transistor action occurs; the emitter current is 0. Thus

$$I_{IH} \approx 0 .$$

Example 5.12

Estimate P_H for the Schottky TTL gate of Figure 5.48.

Solution. With logic-zero input applied from a similar gate (and therefore with the output high), Q_I is on hard; Q_S, Q_O, and Q_D are cut off and Q_P and Q_{P2} are forward active. The DC power dissipation under these conditions is

$$P_H = V_{CC} I_{CCH} = V_{CC}\left(I_{RB} + I_{RC} + I_{RCP}\right).$$

Because $I_{IH} = 0$, the Q_{P2} emitter current is 0 and thus

$$I_{RC} + I_{RCP} = I_{REP} = \frac{V_{CC} - V_{BEA}}{R_C / \left(\beta_F + 1\right) + R_{EP}};$$

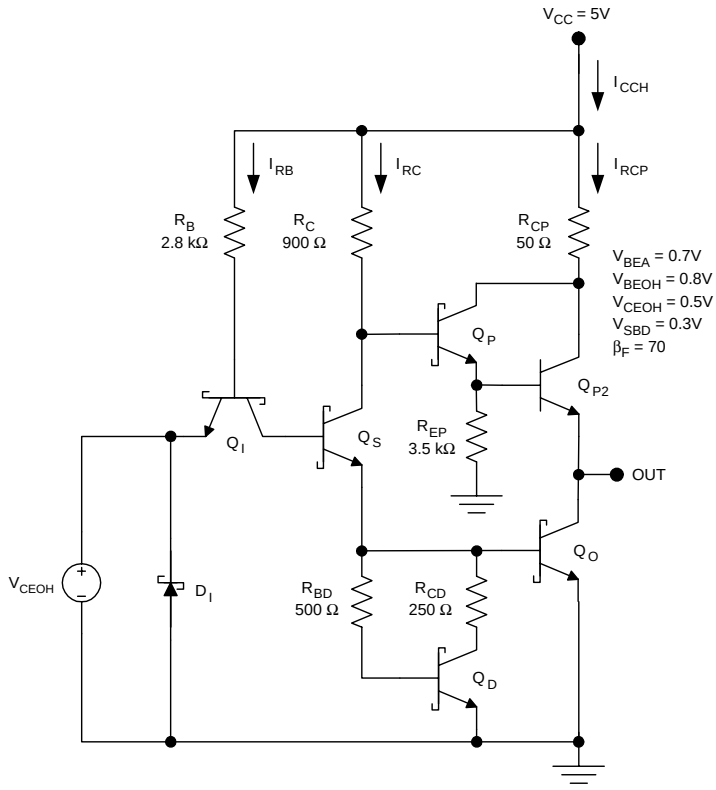


FIGURE 5.48
STTL inverter for the calculation of P_H.

therefore, the output high power is

$$P_H = V_{CC}\left(I_{RB} + I_{REP}\right) = V_{CC}\left(\frac{V_{CC} - V_{BEOH} - V_{OL}}{R_B} + \frac{V_{CC} - V_{BEA}}{R_C/\left(\beta_F + 1\right) + R_{EP}}\right)$$

$$= 5\text{ V}\left(\frac{5\text{ V} - 0.8\text{ V} - 0.5\text{ V}}{2.8\text{ k}\Omega} + \frac{5\text{ V} - 0.7\text{ V}}{0.9\text{ k}\Omega/71 + 3.5\text{ k}\Omega}\right) = 12.7\text{ mW}$$

This value is independent of the fan-out because I_{IH} is negligibly small.

Example 5.13

Determine P_L for the Schottky TTL gate of Figure 5.49.

Solution. If a logic-one input is applied from a similar gate as shown in Figure 5.49, then Q_I operates in the reverse Schottky mode, Q_S, Q_P, and Q_D are on hard, Q_P is forward active, and Q_{P2} is cut off. Under these conditions, the steady power dissipation is

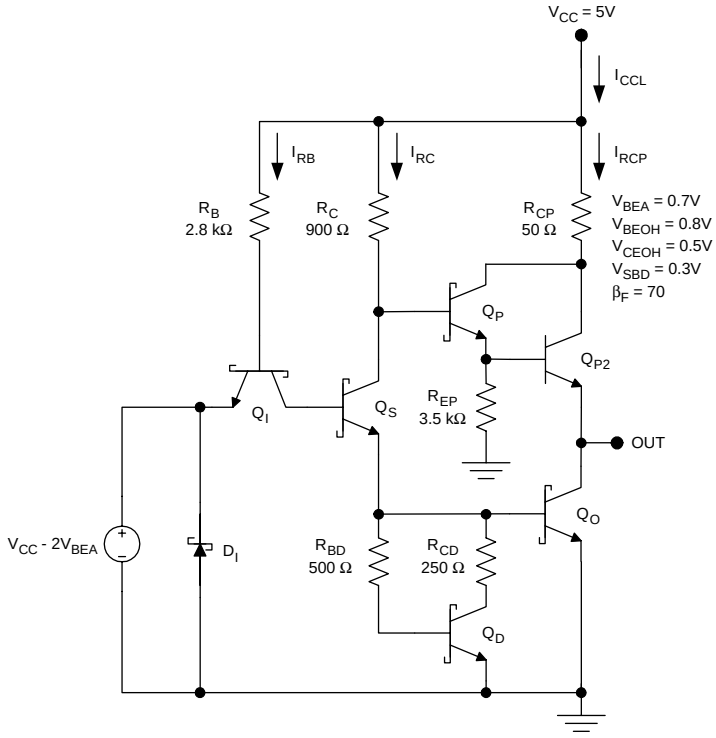


FIGURE 5.49
STTL inverter for the calculation of P_L.

$$P_L = V_{CC} I_{CCL} = V_{CC}\left(I_{RB} + I_{RC} + I_{RCP}\right),$$

where

$$I_{RB} = \frac{V_{CC} - V_{BCOH} - 2V_{BEOH}}{R_B} = \frac{5\text{ V} - 0.3\text{ V} - 1.6\text{ V}}{2.8\text{ k}\Omega} = 1.107\text{ mA},$$

$$I_{RC} = \frac{V_{CC} - V_{CEOH} - V_{BEOH}}{R_C} = \frac{5\text{ V} - 0.5\text{ V} - 0.8\text{ V}}{0.9\text{ k}\Omega} = 4.11\text{ mA}$$

and

$$I_{RCP} \approx I_{REP} = \frac{V_{BEOH} + V_{CEOH} - V_{BEA}}{R_{EP}} = \frac{0.8\text{ V} + 0.5\text{ V} - 0.7\text{ V}}{3.5\text{ k}\Omega} = 0.1714\text{ mA}.$$

Therefore,

$$P_L = 5\text{ V}(1.107\text{ mA} + 4.11\text{ mA} + 0.1714\text{ mA}) = 26.9\text{ mW}.$$

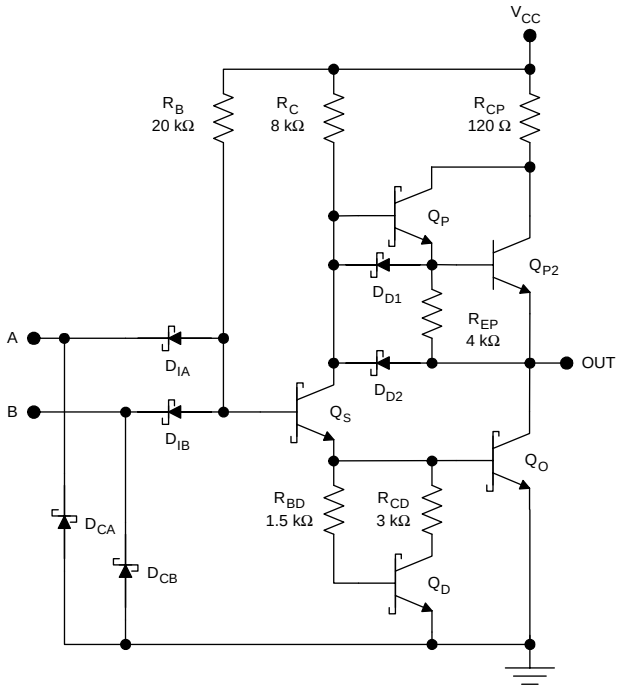


FIGURE 5.50
LSTTL NAND2 gate.

5.9.5 74LS Low-Power Schottky TTL (LSTTL)

The low-power Schottky (74LS/54LS) circuit family evolved from Schottky TTL, with several important changes. These design modifications include 1) scaled resistors, 2) a modified pull-up subcircuit, 3) a DTL input subcircuit, and 4) improved device fabrication. The two-input NAND gate circuit is shown in Figure 5.50.

The resistors were scaled to achieve a reduction in DC dissipation. R_B and R_C, which are most important in determining the dissipation, were scaled up by factors of 7 and 9, respectively. However, other resistors that have less impact on the power dissipation were scaled by smaller factors in order to avoid degradation of the switching speed. The pull-up subcircuit was improved by the inclusion of the two Schottky diodes, D_{D1} and D_{D2}. D_{D1} helps discharge the base of Q_{P2} during high-to-low transitions at the output, thus improving t_{PHL}. D_{D2} helps discharge the load capacitance during high-to-low transitions at the output and also improves t_{PHL}.

The use of Schottky clamping on the drive splitter Q_S obviates the need for a bipolar transistor to remove base charge rapidly from that device. In fact, the Schottky input diodes exhibit smaller parasitic capacitances than the input transistor they replace. Therefore, an improvement in switching speed was achieved by using a DTL input in this TTL circuit.

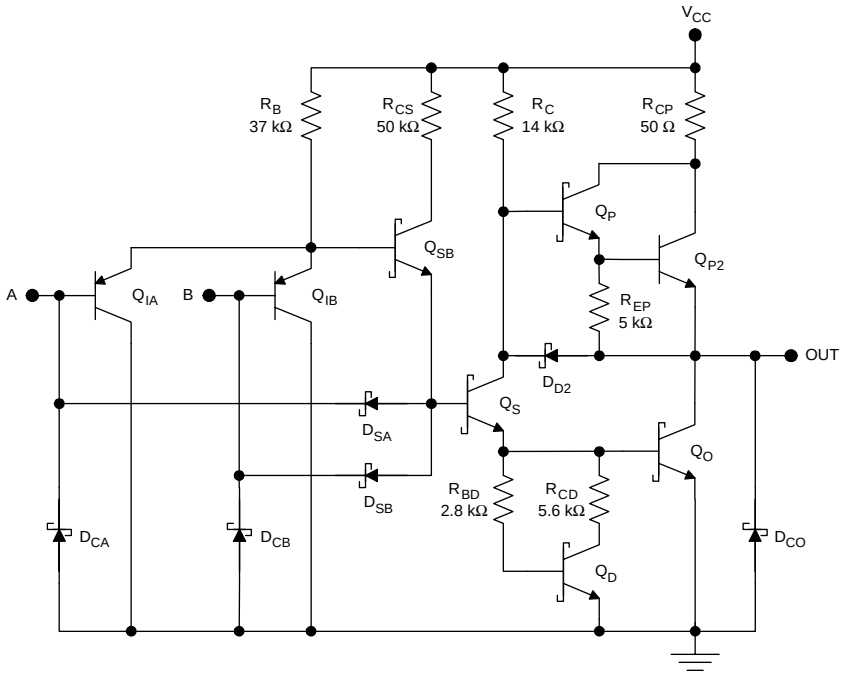


FIGURE 5.51
ALSTTL NAND2 gate.

5.9.6 74ALS Advanced Low-Power Schottky TTL (ALSTTL)

The advanced low-power Schottky TTL (74ALS/54ALS) circuit family shown in Figure 5.51[1] evolved directly from LSTTL and is still popular in modern designs. The key engineering changes compared to LSTTL are 1) replacement of the input diodes with pnp emitter followers; 2) additional npn emitter follower stage Q_{SB}; 3) scaled resistors for reduced power consumption; and 4) improved device fabrication for lower parasitic capacitances.

The pnp input transistors perform the ANDing of the input signals in much the same way as the diodes they replace. The important difference is that the current gain in these pnp emitter followers reduces the input low current and provides superior drive for the drive splitter under switching conditions. Therefore, DC and dynamic performance is improved, despite the fact that the slice fabrication steps are optimized for npn transistors; thus, lateral pnps with modest current gain are used here.

The added npn emitter follower Q_{SB} benefits the DC and dynamic performance of the circuit. First, the base emitter drop of this transistor compensates for the nearly equal but opposite base emitter drop in the pnp. The end result is that V_{IL} and V_{IH} are increased compared to LSTTL, providing an improved low noise margin and a better balance between the two noise margins. Second, the added emitter follower provides more drive current to

the drive splitter Q_S under dynamic conditions. Thus, the high-to-low propagation delay is improved.

The transistors utilized are scaled down in size and use oxide isolation rather than junction isolation, for smaller parasitic capacitances and smaller parasitic resistances. The resulting devices have superior gain bandwidth products. The use of scaled up resistors reduces the DC dissipation in ALSTTL, in a direct trade-off with switching speed. However, the other circuit and device improvements described previously more than compensate for the scaled resistors, so propagation delays are significantly improved over LSTTL. The combination of these factors results in performance (P_{DC} = 1 mW, t_P = 4 ns @ 15 pF, PDP = 4 pJ @ 15 pF) that is entirely acceptable for modern applications on motherboards of workstations and servers.

5.9.7 74F Fairchild Advanced Schottky TTL (FAST)

The Fairchild advanced Schottky TTL (74F/54F) family illustrated in Figure 5.52,[2] sometimes called "FAST," also evolved from LSTTL but with a different mix of performance compared to ALSTTL. Here greater emphasis has been placed on speed than on power, so the resistors were scaled down significantly. The transistor fabrication is greatly improved over the LSTTL and the oxide isolated devices are comparable to those used in ALSTTL. There are two important differences in circuit approach compared to ALSTTL: 1) the DTL input is retained, thus obviating the need for lateral pnp fabrication and 2) a "Miller killer" subcircuit has been added for improved low-to-high dynamic response.

The use of Schottky diodes rather than pnp input transistors results in lower parasitics and faster switching for the input devices. The penalty is increased I_{IL}, but this is acceptable in FAST. The across-the-board downscaling of resistors allows a larger value of I_{IL} without a degradation in the DC fan-out by the same factor.

The Miller killer subcircuit, comprising D_V, Q_K, D_{BK}, and D_{CK}, improves the dynamic performance of the circuit. D_V is a varactor diode that is always reverse biased. The term "varactor" comes from the term "variable reactor" and signifies that this diode is used as a voltage variable capacitance. Under DC conditions, the varactor does not conduct and so has no effect on the circuit. During an output low-to-high transition, Q_S turns off before Q_O. When the voltage begins to rise at the collector of Q_S, D_V conducts a displacement current to the base of the killer transistor Q_K. This transistor is briefly driven in the on-hard mode and can sink a large amount of current from the base of the output transistor Q_O. The greatly improved turn-off base current for the output transistor provides a drastically shortened t_{PLH}. The colorful name for this subcircuit stems from the fact that it mitigates the Miller effect that is the dominant limitation on the low-to-high propagation delay.

Compared to ALSTTL, the FAST circuit offerings achieve faster switching (t_P = 2.5 ns @ 15 pF), but with higher power dissipation (P_{DC} = 4 mW). The

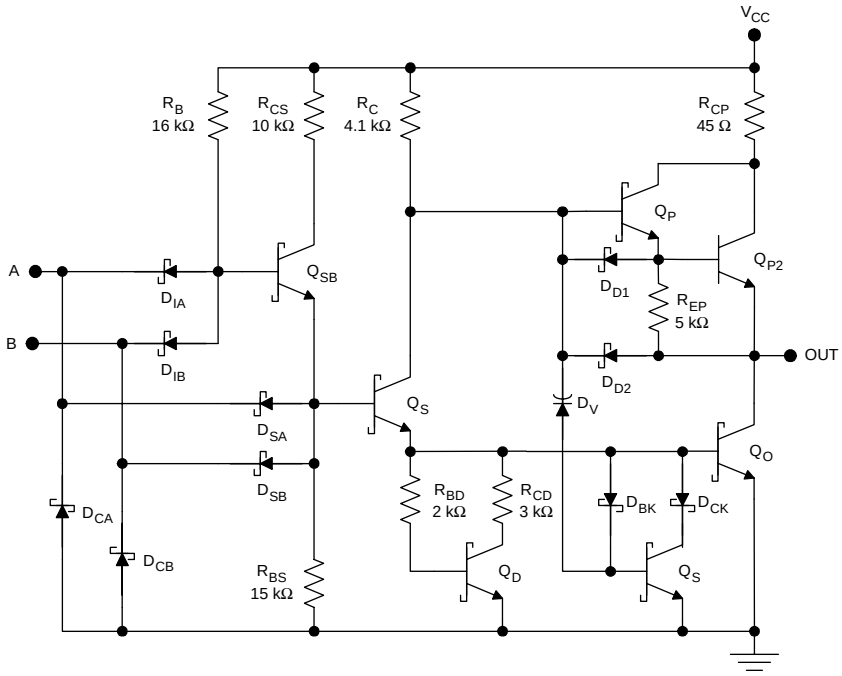


FIGURE 5.52
FAST NAND2 gate.

overall power delay product (10 pJ) is inferior to that of ALSTTL. Nonetheless, FAST gates are in common use today in SSI and MSI applications requiring high off-chip data rates, such as bus-organized motherboards. In these applications, speed is more important than power and the improvement in switching speed translates directly to higher bit rates.

5.9.8 74AS Advanced Schottky TTL (ASTTL)

The advanced Schottky TTL (74AS/54AS) family shown in Figure 5.53[1] evolved concurrently with the ALSTTL and FAST families and with equally advanced fabrication. Therefore, the transistors have equivalent gain bandwidth products in all three families and the performance differences can be ascribed totally to circuit engineering. Notable aspects of the circuit design include: 1) the resistors have been down-scaled to improve the speed in a trade-off with power; 2) input diodes and pnp input transistors have been used together; 3) the input clamping diodes have been replaced by clamp transistors; 4) the discharge diode between the output node and the collector of the drive splitter has been replaced by a discharge transistor Q_{OD}; 5) a Miller killer subcircuit has been included; 6) a low impedance driver, Q_{P3}, has been added to the pull-up subcircuit; and 7) output clamping diodes have been inserted.

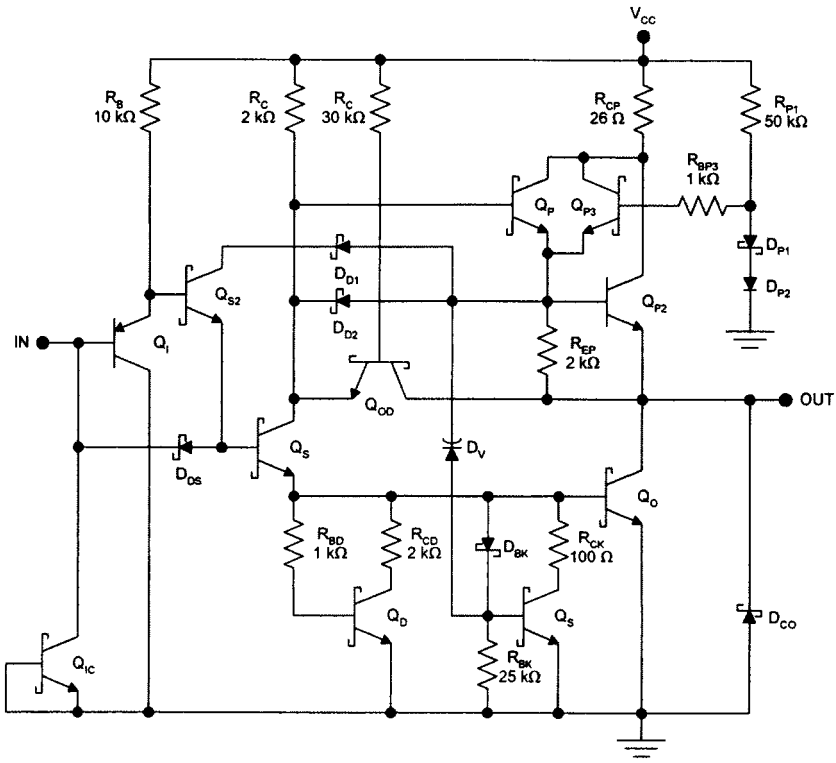


FIGURE 5.53
ASTTL inverter.

The ASTTL family of products brings a mix of performance attributes emphasizing speed over low power. The propagation delays (t_P = 1.5 ns @ 15 pF) are the shortest available in any TTL family but come at the cost of a fivefold increase in dissipation compared to FAST (20 mW vs. 4 mW). These circuits are the best TTL choice for SSI and MSI applications requiring high data rates on bus-organized motherboards.

5.10 PSPICE Simulations: BJT Inverter

Simulations were performed using PSPICE 9.1, which can be downloaded free.[3] The BJT model parameters used in all simulations are provided in Table 5.1.

5.10.1 Voltage Transfer Characteristic

The voltage transfer characteristic was simulated for the *unloaded* RTL inverter of Figure 5.54 with V_{CC} = 5 V. The results are shown in Figure 5.55. The critical voltages are: V_{OH} = 5 V, V_{OL} = 0.056 V, V_{IL} = 0.61 V and V_{IH} = 1.26 V.

TABLE 5.1

BJT SPICE Parameters

Parameter	Value	Units
IS	2.0f	A
BF	70	—
NF	1	—
BR	0.5	—
NR	1	—
CJE	0.3p	F
VJE	0.8	V
MJE	0.333	—
TF	0.2n	s
CJC	0.15p	F
VJC	0.75	V
MJC	0.5	—
TR	10n	s

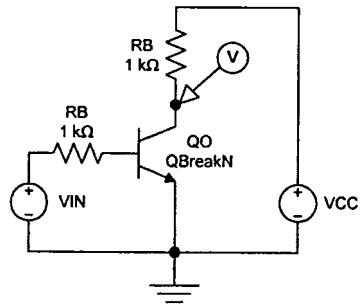


FIGURE 5.54
RTL circuit used for the simulation of the VTC.

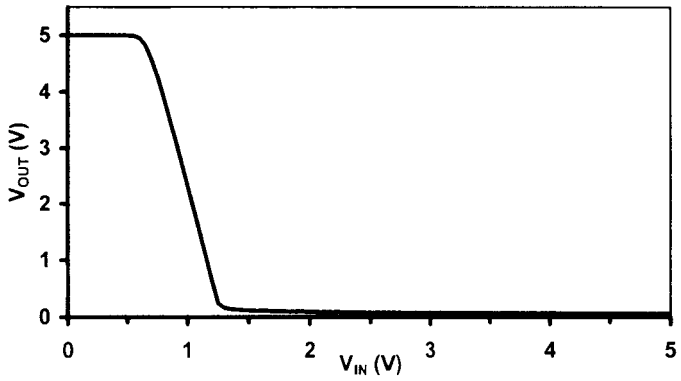


FIGURE 5.55
Simulated VTC for the RTL circuit.

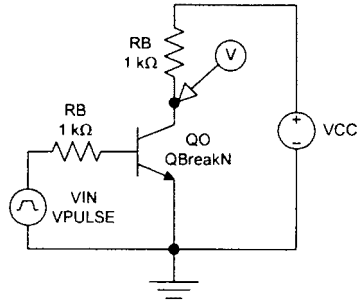


FIGURE 5.56
Unloaded RTL circuit for the simulation of the transient response.

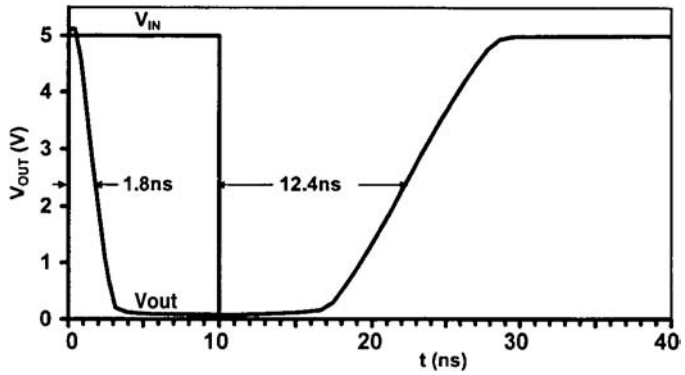


FIGURE 5.57
Simulated RTL transient response.

5.10.2 Propagation Delays

The propagation delays for the *unloaded* RTL inverter were determined using the circuit of Figure 5.56. The pulse source parameters were V1 = 5 V, V2 = 0, TD = TR = TF = 0, PW = 10 ns, and PER = 40 ns. The results of the transient simulation appear in Figure 5.57. The propagation delays can be determined using the 50% points on the input and output waveforms and are t_{PHL} = 1.8 ns and t_{PLH} = 12.4 ns.

5.11 PSPICE Simulations: TTL

Simulations were performed using PSPICE 9.1, which can be downloaded free.[3] The bipolar device parameters used in all simulations are provided in Table 5.1 and Table 5.2.

TABLE 5.2

Diode SPICE Parameters

Parameter	Value	Units
IS	2.0f	A
N	1	—
TT	0.2n	s
CJO	0.3p	F
VJ	0.8	V
M	0.5	—

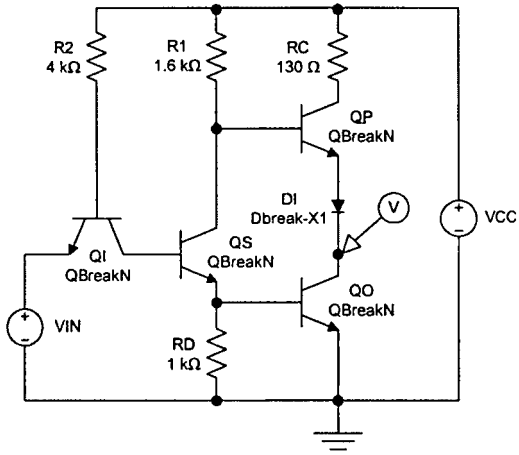


FIGURE 5.58
TTL circuit used for the simulation of the VTCs.

5.11.1 Voltage Transfer Characteristic

The voltage transfer characteristic was simulated for the *unloaded* TTL inverter of Figure 5.58 with V_{CC} as a parameter. The results of Figure 5.59 show that, apart from V_{OH}, the critical voltages are unaffected by the value of the supply voltage.

5.11.2 DC Dissipation

The DC dissipation was analyzed using a DC sweep and the circuit of Figure 5.60. The I_{CC} vs. V_{IN} characteristic appears in Figure 5.61. The results show that I_{CCH} = 1.05 mA (P_H = 5.25 mW) and I_{CCL} = 3.35 mA (P_L = 16.75 mW).

5.11.3 Input Current

The DC input was analyzed using a DC sweep and the circuit of Figure 5.62. The I_{IN} vs. V_{IN} characteristic appears in Figure 5.63. The results show that I_{IL} = 1.05 mA (@ VIN = 0.1 V) and I_{IH} = 0.36 mA (@ VIN = 3.4 V).

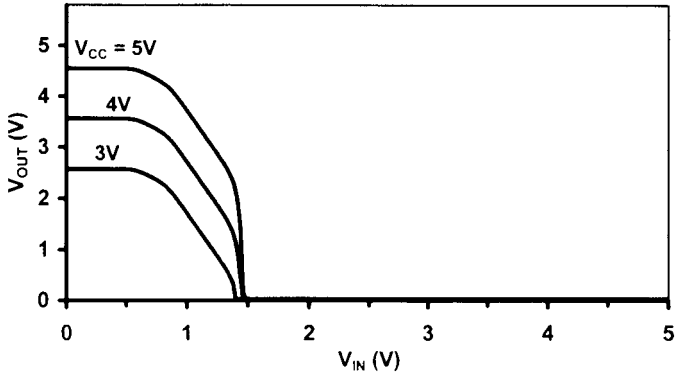


FIGURE 5.59
Simulated VTCs for the TTL circuit.

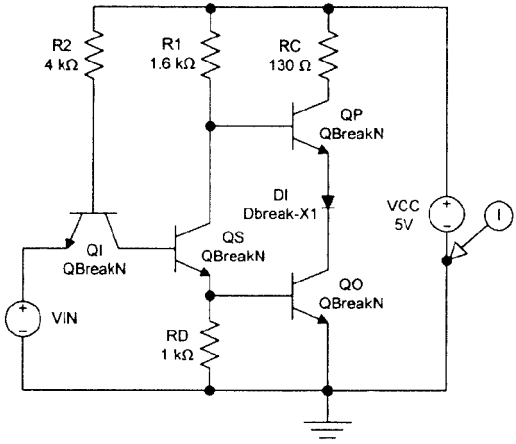


FIGURE 5.60
TTL circuit for the determination of P_L and P_H.

5.11.4 Output Current

The circuit of Figure 5.64 was used with a DC sweep to determine I_{OL} for the TTL circuit. The V_{OUT} vs. I_{OUT} characteristic appears in Figure 5.65. The results show that the output voltage rises sharply for $I_{OUT} > 190$ mA, indicating that the output transistor has come out of saturation. Based on this observation, $I_{OL} = 190$ mA.

5.11.5 Propagation Delays

The propagation delays for a TTL inverter with a 15-pF load were determined using the circuit of Figure 5.66. The pulse source parameters were

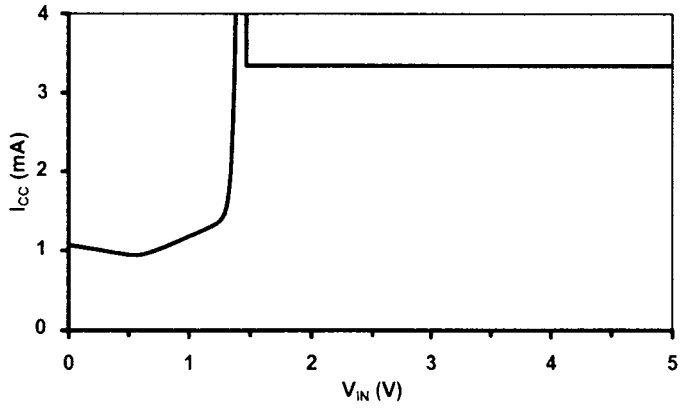


FIGURE 5.61
Simulated I_{CC} vs. V_{IN} characteristic for the TTL circuit.

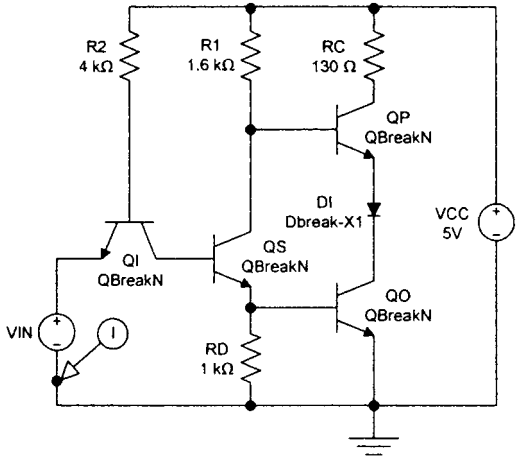


FIGURE 5.62
TTL circuit for the determination of I_{IL} and I_{IH}.

V1 = 5 V, V2 = 0, TD = TR = TF = 0, PW = 20 ns, and PER = 40 ns. The results of the transient simulation appear in Figure 5.67. As a consequence of the saturation delay for the output transistor, t_{PLH} is considerably longer than t_{PHL}.

5.12 PSPICE Simulations: LSTTL

Simulations were performed using PSPICE 9.1, which can be downloaded free.[3] The BJT model parameters used in all simulations were the same as those used for the TTL simulations. Schottky diodes were modeled using

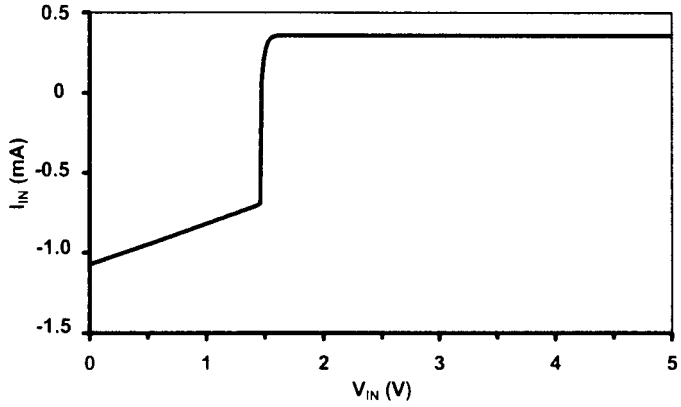


FIGURE 5.63
Simulated I_{IN} vs. V_{IN} characteristic for the TTL circuit.

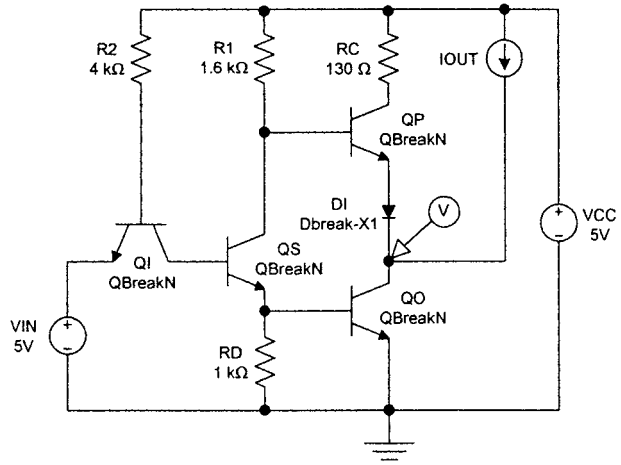


FIGURE 5.64
TTL circuit for the determination of I_{OL}.

the parameters in Table 5.3. SPICE simulators do not include Schottky-clamped bipolar transistors. Therefore, each clamping diode must be included explicitly in the schematic.

5.12.1 Voltage Transfer Characteristic

The voltage transfer characteristic was simulated for the *unloaded* LSTTL inverter of Figure 5.68 with V_{CC} as a parameter. A nested sweep was used to vary V_{CC}. The results of Figure 5.69 show that the voltage transfer characteristics are "squared" by the inclusion of Q_D and that the Schottky clamping increases V_{OL}.

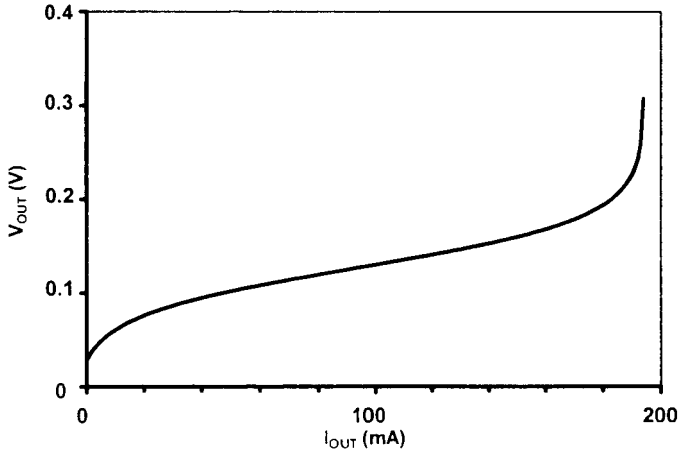


FIGURE 5.65
Simulated I_{CC} vs. V_{IN} characteristic for the TTL circuit.

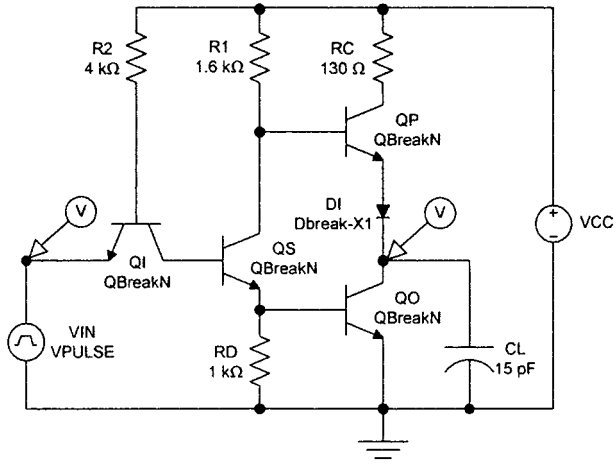


FIGURE 5.66
TTL circuit for the simulation of the propagation delays with a 15-pF load.

5.12.2 Propagation Delays

The propagation delays for an LSTTL inverter with a 15-pF load were determined using the circuit of Figure 5.70. The pulse source parameters were V1 = 5 V, V2 = 0, TD = TR = TF = 0, PW = 20 ns, and PER = 40 ns. The results of the transient simulation appear in Figure 5.71. Both propagation delays have shortened considerably compared to the case of TTL. Additional transient simulations were done to determine the propagation delays as a function of the load capacitance; the results are shown in Figure 5.72. Both

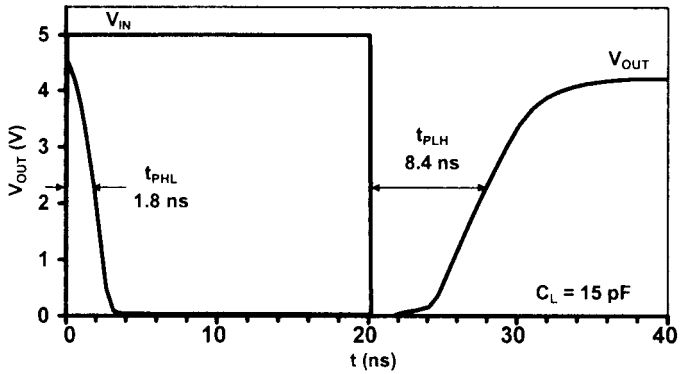


FIGURE 5.67
Simulated TTL transient response with a 15-pF load.

TABLE 5.3

Schottky Diode SPICE Parameters for LSTTL Simulations

Parameter	Value	Units
IS	9.0n	A
N	1	—
TT	0	s
CJO	0.05p	F
VJ	0.4	V
M	0.5	—

propagation delays increase monotonically with the load capacitance and, moreover, the characteristics are both approximately linear with slopes of about 100 Ω.

5.13 Summary

Transistor–transistor logic (TTL) is a family of saturated bipolar logic gates. TTL circuits are important today because of their ability to drive large load capacitances at relatively high data rates. As with other bipolar logic circuits, the dissipation and fan-out of TTL gates are determined by DC considerations.

TTL gates provide good dynamic response and are relatively immune to the effects of capacitive loads because of the use of active pull-up and pull-down circuitry in a so-called "totem-pole output." The primary limitation on the switching speed of standard TTL is the saturation delay of the output pull-down transistor.

Modern versions of TTL use Schottky clamped transistors to prevent hard saturation, as well as squaring circuits and pseudo Darlington pull-up circuitry.

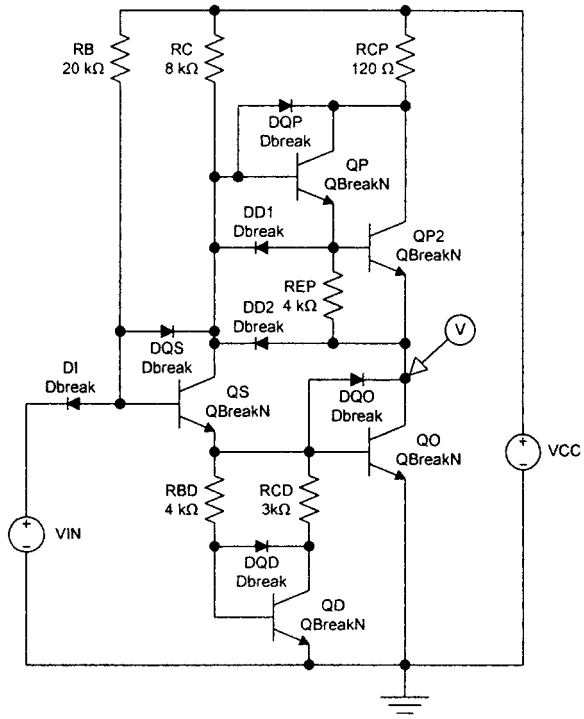


FIGURE 5.68
LSTTL circuit used for the simulation of the VTCs.

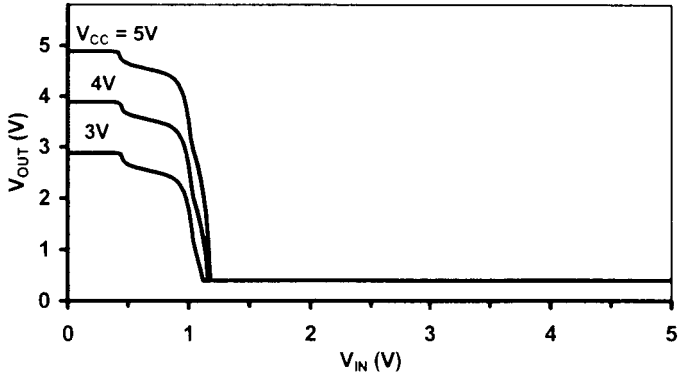


FIGURE 5.69
Simulated VTCs for the LSTTL circuit.

These modifications, plus fabrication improvements in the bipolar transistors, have resulted in shorter propagation delays. The modern TTL families of commercial importance include advanced Schottky TTL (ASTTL), Fairchild advanced Schottky TTL (FAST), and advanced low-power Schottky TTL (ALSTTL).

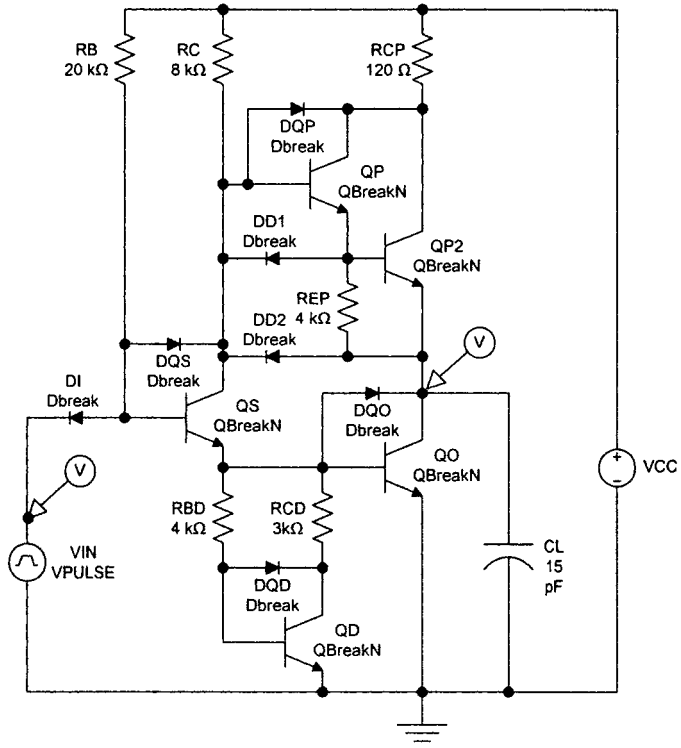


FIGURE 5.70
LSTTL circuit for the simulation of the propagation delays with a 15-pF load.

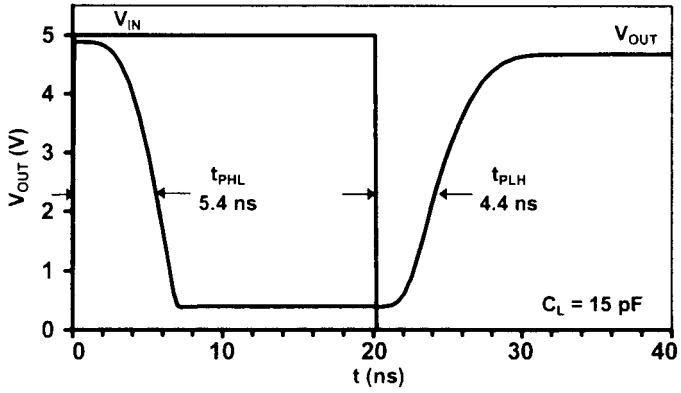


FIGURE 5.71
Simulated LSTTL transient response with a 15-pF load.

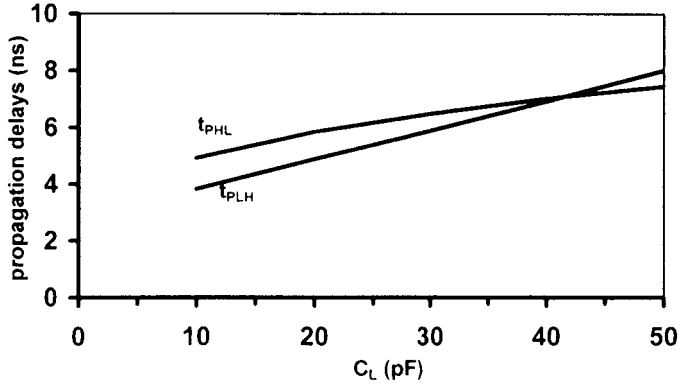
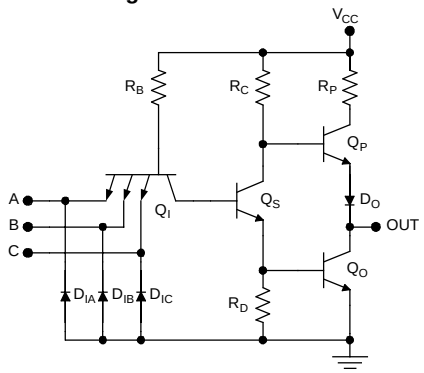


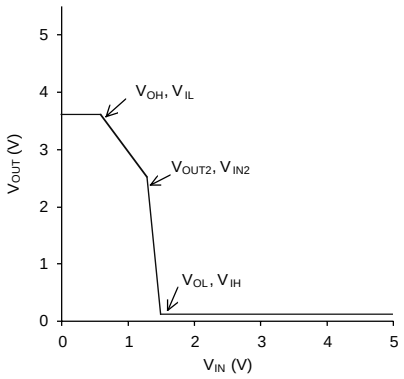
FIGURE 5.72
LSTTL propagation delays as a function of the load capacitance.

TRANSISTOR–TRANSISTOR LOGIC QUICK REFERENCE

TTL NAND3 gate circuit



DC Voltage Transfer Characteristic



Transistor-transistor logic (TTL) uses bipolar transistors in the pull-up and pull-down circuits, allowing high off-chip data rates with large capacitive loads. Schottky clamped TTL(LSTTL, ALSTL, ASTTL, and FAST) avoids saturation and achieves much shorter propagation delays.

DC Voltage Transfer Characteristic

First Breakpoint $V_{IL} = V_{BEA} - V_{CES}$ $\quad V_{OH} = V_{CC} - V_{BEA} - V_D$

Second Breakpoint $V_{IN2} = 2V_{BEA} - V_{CES}$ $\quad V_{OUT2} = V_{CC} - R_C\left(\frac{V_{BEA}}{R_D}\right) - V_{BEA} - V_D$

Third Breakpoint $V_{IH} = 2V_{BES} - V_{CES}$ $\quad V_{OL} = V_{CES}$

DC Output Current

$$I_{OL} = \sigma_{MAX}\beta_F\left[(\beta_R + 1)\left(\frac{V_{CC} - V_{BCA} - 2V_{BES}}{R_B}\right) + \frac{V_{CC} - V_{CES} - V_{BES}}{R_C} - \frac{V_{BES}}{R_D}\right]$$

DC Input Current

$$I_{IL} = \frac{V_{CC} - V_{BES} - V_{CES}}{R_B} \qquad I_{IH} = \frac{\beta_R}{k}\left(\frac{V_{CC} - V_{BCA} - 2V_{BES}}{R_B}\right)$$

DC Fan-out

$$N_{MAX} \le \frac{I_{OL}}{I_{IL}}$$

DC Dissipation

$$P_L = V_{CC}\left(\frac{V_{CC} - V_{BCA} - 2V_{BES}}{R_B} + \frac{V_{CC} - V_{CES} - V_{BES}}{R_C}\right)$$

$$P_H = V_{CC}\left(\frac{V_{CC} - V_{BES} - V_{CES}}{R_B} + N\beta_R\left(\frac{V_{CC} - V_{BCA} - 2V_{BES}}{R_B}\right)\right)$$

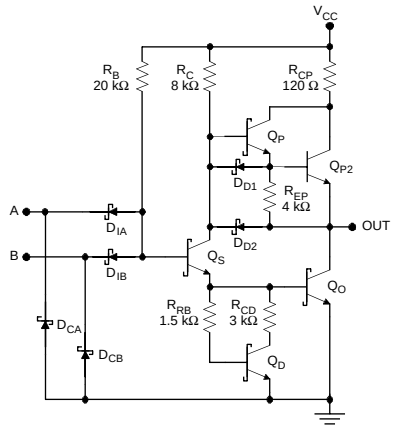
Propagation Delays

$t_{PLH} > t_{PHL}$, limited by the saturation delay for the output transistor. Schottky clamping greatly reduces the propagation delays in all modern TTL families.

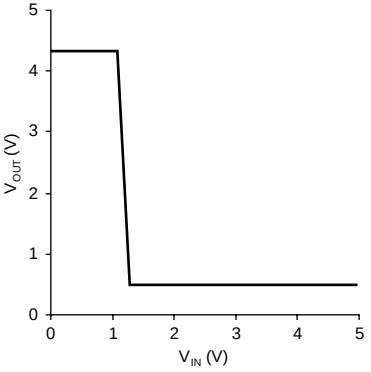
f = 10^{-15} p = 10^{-12} n = 10^{-9} μ = 10^{-6} m = 10^{-3} k = 10^{3} M = 10^{6} G = 10^{9}

LOW-POWER SCHOTTKY TRANSISTOR–TRANSISTOR LOGIC QUICK REFERENCE

LSTTL NAND2 gate circuit



DC Voltage Transfer Characteristic



In Low-Power Schottky TTL, Schottky diodes are used to clamp the bipolar transistors and avoid hard saturation. The commercially important families of TTL (ALSTL, ASTTL, and FAST) are all high-performance derivatives of LSTTL.

DC Voltage Transfer Characteristic

$V_{IL} = 2V_{BEA} - V_D$ $\qquad$ $V_{IH} = 2V_{BE}(\text{on hard}) - V_D$

$V_{OL} = V_{CE}(\text{on hard})$ $\qquad$ $V_{OH} = V_{CC} - V_{BEA}$

DC Output Current

$$I_{OL} = \sigma_{MAX}\beta_F\left[\left(\frac{V_{CC} - 2V_{BEOH}}{R_B}\right) + \frac{V_{CC} - V_{CEOH} - V_{BEOH}}{R_C} - \frac{V_{BEOH} - V_{CEOH}}{R_{CD}}\right]$$

DC Input Current

$$I_{IL} = \frac{V_{CC} - V_D - V_{CEOH}}{R_B} \qquad I_{IH} \approx 0$$

DC Fan-out

$$N_{MAX} \le \frac{I_{OL}}{I_{IL}}$$

DC Dissipation

$$P_L = V_{CC}\left(\frac{V_{CC} - 2V_{BEOH}}{R_B} + \frac{V_{CC} - V_{CEOH} - V_{BEOH}}{R_C} + \frac{V_{BEOH} + V_{BEA}}{R_{EP}}\right)$$

$$P_H = V_{CC}\left(\frac{V_{CC} - V_D - V_{CEOH}}{R_B}\right)$$

Propagation Delays

Schottky TTL gates exhibit short propagation delays because of the avoidance of hard saturation in the bipolar transistors. They are also relatively insensitive to capacitive loading because of the use of a totem-pole output with a pseudo-Darlington pull-up circuit.

f = 10^{-15} p = 10^{-12} n = 10^{-9} μ = 10^{-6} m = 10^{-3} k = 10^{3} M = 10^{6} G = 10^{9}

Laboratory Exercises

L5.1. Consider the RTL inverter of Figure 5.73 with a *lumped capacitive load*.

1. Using hand calculations, determine and plot the average propagation delay vs. the load capacitance.
2. Using SPICE transient simulations for a single RTL inverter circuit, determine and plot the average propagation delay vs. the load capacitance.
3. Build a three-stage ring oscillator using the RTL circuit. By varying the load capacitance (applied at each stage), determine and plot the average propagation delay vs. the load capacitance. Is the amplitude of the oscillations equal to the expected logic swing, $V_{OH} - V_{OL}$?
4. Plot the hand-calculated, simulated, and experimental results together for direct comparison. Determine the slopes of the characteristics, in ohms. Are there significant differences (>10%) between the sets of data?
5. Use SPICE to predict the ring oscillator frequency vs. the load capacitance. Are these results in agreement with the transient results for a single RTL gate? Are they in agreement with the experimental results?

L5.2. Consider the RTL inverter of Figure 5.74 with *scaled resistors* and a *lumped capacitive load*.

1. Using SPICE, determine and plot P_{DC} vs. x, t_P vs. x, and the PDP vs. x. Is there a value of x that minimizes the power delay product?
2. Using experimental measurements of P_H and P_L, calculate P_{DC} as a function of x.
3. Using three-stage ring oscillators, determine the average propagation delay as a function of x.
4. Plot the PDP vs. x (both simulated and experimental results).

L5.3. Consider the standard TTL gate circuit of Figure 5.75.

1. Determine and plot the voltage transfer characteristic using hand calculations.
2. Determine and plot the voltage transfer characteristic using SPICE.
3. Estimate P_H, P_L, and P_{DC} by hand calculations.
4. Determine P_H, P_L, and P_{DC} using SPICE.
5. Estimate the t_{PLH} by hand calculations.
6. Estimate the propagation delays using SPICE.

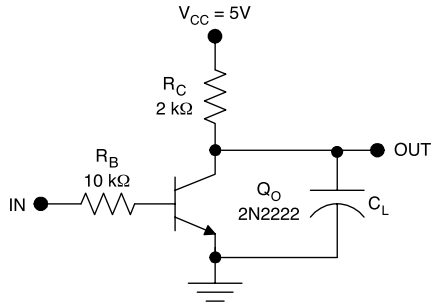


FIGURE 5.73
RTL inverter (L5.1).

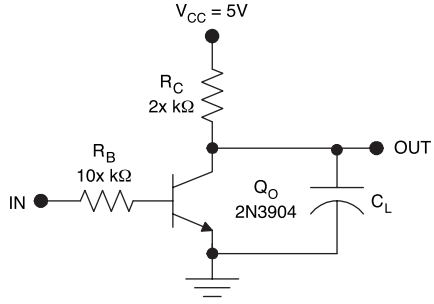


FIGURE 5.74
RTL inverter (L5.2).

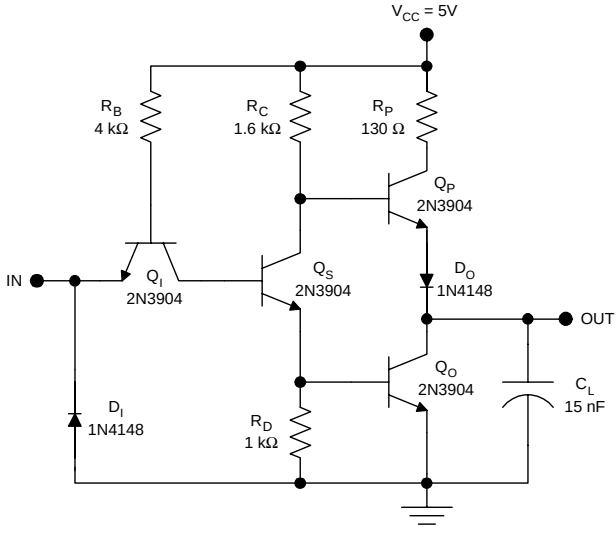


FIGURE 5.75
Standard TTL gate circuit (L5.3).

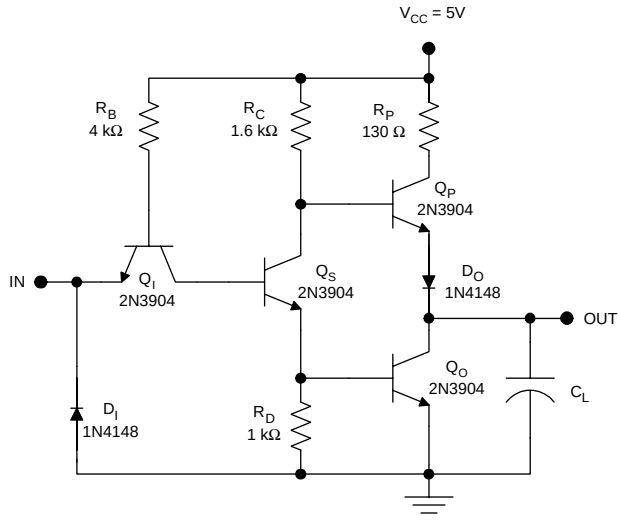


FIGURE 5.76
TTL inverter (L5.4).

7. Build the inverter and measure the voltage transfer characteristic using the x–y feature of an oscilloscope or virtual instrument and a low frequency input signal (1 kHz).
8. Build a seven-stage ring oscillator using seven of these inverter circuits; measure the frequency of oscillation and determine the average propagation delay from the measured frequency of oscillation.

L5.4. Consider the TTL inverter of Figure 5.76 with a *lumped capacitive load*. 5 nF < C_L < 100 nF.

1. Determine and plot the voltage transfer characteristic using hand calculations.
2. Determine and plot the voltage transfer characteristic using SPICE.
3. Estimate P_H, P_L, and P_{DC} by hand calculations.
4. Determine P_H, P_L, and P_{DC} using SPICE.
5. Estimate the propagation delays by hand calculations and plot the average propagation delay as a function of C_L.
6. Estimate the propagation delays using SPICE and plot the average propagation delay as a function of C_L.
7. Build the inverter and measure the voltage transfer characteristic using the x–y feature of an oscilloscope or virtual instrument and a low frequency input signal (1 kHz).
8. Using seven-stage ring oscillators, determine the average propagation delay as a function of the load capacitance. Plot t_p vs. C_L.

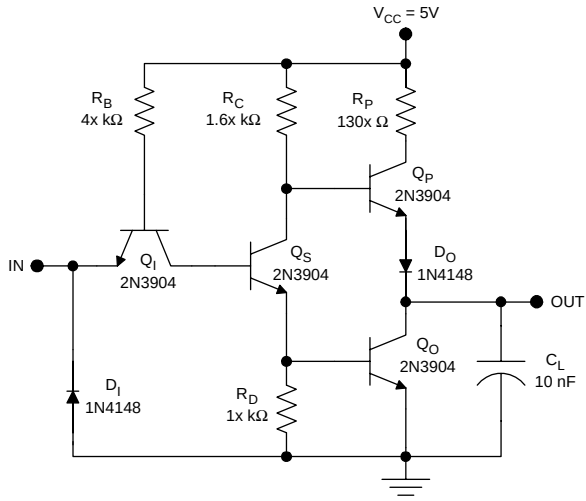


FIGURE 5.77
TTL inverter (L5.5).

L5.5. Consider the TTL inverter of Figure 5.77 with *scaled resistors*. $1/4 < x < 4$.

1. Estimate P_H, P_L, and P_{DC} by hand calculations. Plot P_{DC} vs. x.
2. Determine P_H, P_L, and P_{DC} using SPICE. Plot P_{DC} vs. x.
3. Estimate the propagation delays using SPICE and plot the average propagation delay as a function of x.
4. Using the SPICE results, plot the power delay product vs. x.
5. Build the inverter and measure P_H, P_L, and P_{DC} as a function of x.
6. Using seven-stage ring oscillators, determine the average propagation delay as a function of x.
7. From the measured results, plot the power delay product vs. x and compare to the SPICE results.

L5.6. Consider the TTL inverter of Figure 5.78. 4 V < V_{CC} < 10 V.

1. Determine and plot the voltage transfer characteristic using hand calculations. Plot V_{OUT} vs. V_{IN} with V_{CC} as a parameter.
2. Determine and plot the voltage transfer characteristic using SPICE. Plot V_{OUT} vs. V_{IN} with V_{CC} as a parameter.
3. Estimate P_H, P_L, and P_{DC} by hand calculations. Plot P_{DC} vs. V_{CC}.
4. Determine P_H, P_L, and P_{DC} using SPICE. Plot P_{DC} vs. V_{CC}.
5. Estimate the propagation delays using SPICE and plot the average propagation delay as a function of V_{CC}.
6. Build the inverter and measure the voltage transfer characteristic using the x–y feature of an oscilloscope or virtual instrument

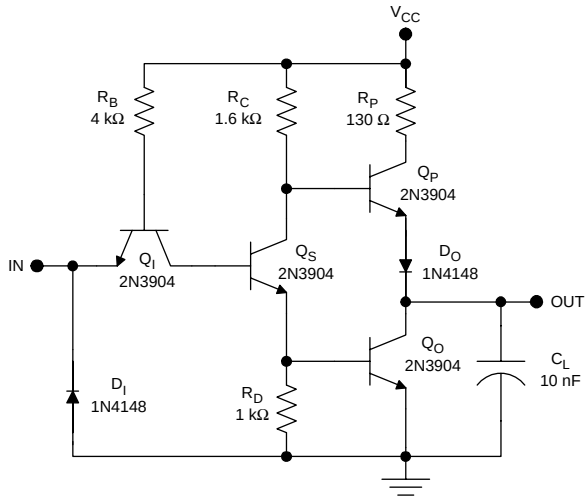


FIGURE 5.78
TTL inverter (L5.6).

and a low frequency input signal (1 kHz). Plot V_{OUT} vs. V_{IN} with V_{CC} as a parameter.

7. Using seven-stage ring oscillators, determine the average propagation delay as a function of V_{CC}.

L5.7. Consider the TTL inverter of Figure 5.79 with a *lumped capacitive load* and *variable supply voltage*. 5 nF < C_L < 100 nF and 4 V < V_{CC} < 10 V.

1. Determine and plot the voltage transfer characteristic using hand calculations. Plot V_{OUT} vs. V_{IN} with V_{CC} as a parameter.
2. Determine and plot the voltage transfer characteristic using SPICE. Plot V_{OUT} vs. V_{IN} with V_{CC} as a parameter.
3. Estimate P_H, P_L, and P_{DC} by hand calculations. Plot P_{DC} vs. V_{CC}.
4. Determine P_H, P_L, and P_{DC} using SPICE. Plot P_{DC} vs. V_{CC}.
5. Estimate the propagation delays using SPICE and plot the average propagation delay as a function of C_L with V_{CC} as a parameter.
6. Build the inverter and measure the voltage transfer characteristic using the x–y feature of an oscilloscope or virtual instrument and a low frequency input signal (1 kHz). Plot V_{OUT} vs. V_{IN} with V_{CC} as a parameter.
7. Using three-stage ring oscillators, determine the average propagation delay as a function of the load capacitance. Plot t_P vs. C_L with V_{CC} as a parameter.

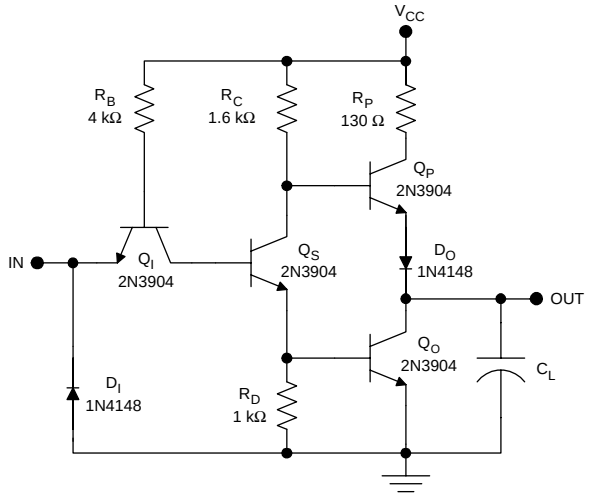


FIGURE 5.79
TTL inverter (L5.7).

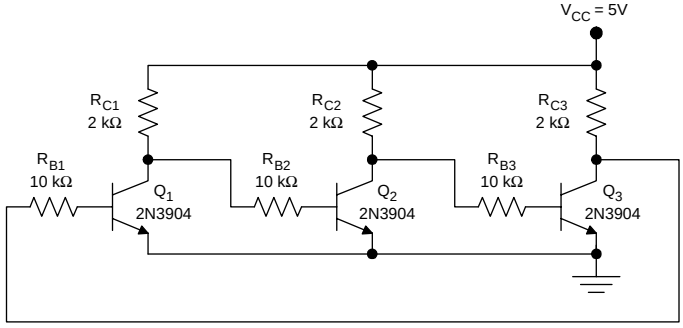


FIGURE 5.80
Bipolar transistor ring oscillators without clamping (L5.8).

L5.8. Consider the bipolar transistor ring oscillators without clamping (Figure 5.80) and with Schottky clamping (Figure 5.81).

1. Using hand calculations, estimate the frequencies of oscillation for the unclamped and clamped ring oscillators.
2. Using SPICE, estimate the frequencies of oscillation for the unclamped and clamped ring oscillators.
3. Build each of the ring oscillators and measure the frequencies of oscillation. Also determine the peak-to-peak amplitude of the oscillations in each case.
4. Compare the hand-calculated, simulated, and experimental results. Contrast the results for the clamped and unclamped

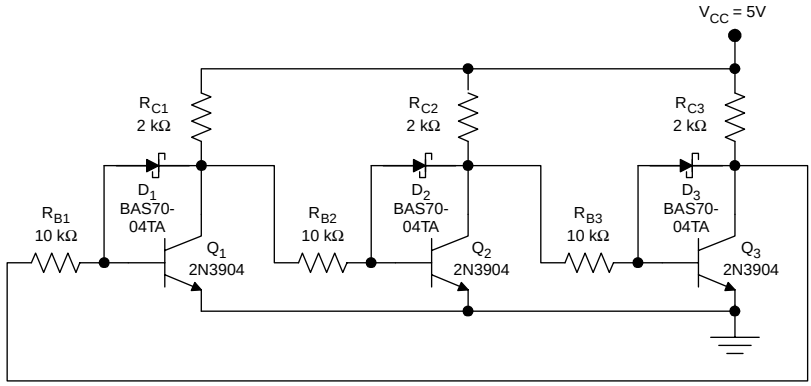


FIGURE 5.81
Bipolar transistor ring oscillator with Schottky clamping (L5.8).

circuits. Are the discrepancies between the hand-calculated and experimental results related to the different peak-to-peak amplitudes?

L5.9. Obtain a 74F04 hex inverter integrated circuit.

1. Determine the voltage transfer characteristic using hand calculations and SPICE, and by experiment with V_{CC} = 5 V. Plot all three characteristics on one graph for comparison.
2. Estimate the propagation delays using SPICE and plot the average propagation delay as a function of C_L with V_{CC} as a parameter. 10 pF < C_L < 100 pF and 4.5 V < V_{CC} < 5.5 V.
3. Experimentally determine the propagation delays using five-stage ring oscillators. Plot the experimental and SPICE results together for direct comparison. (Plot the average propagation delay as a function of C_L with V_{CC} as a parameter. 10 pF < C_L < 100 pF and 4.5 V < V_{CC} < 5.5 V.) What is the significance of the slopes of the characteristics?

L5.10. Obtain a 74AS04 hex inverter integrated circuit.

1. Determine the voltage transfer characteristic using hand calculations and SPICE, and by experiment with V_{CC} = 5 V. Plot all three characteristics on one graph for comparison.
2. Estimate the propagation delays using SPICE and plot the average propagation delay as a function of C_L with V_{CC} as a parameter. 10 pF < C_L < 100 pF and 4.5 V < V_{CC} < 5.5 V.
3. Experimentally determine the propagation delays using five-stage ring oscillators. Plot the experimental and SPICE results together for direct comparison. (Plot the average propagation delay as a function of C_L with V_{CC} as a parameter. 10 pF < C_L < 100 pF and 4.5 V < V_{CC} < 5.5 V.) What is the significance of the slopes of the characteristics?

Problems

P5.1. For the *unloaded* RTL inverter of Figure 5.82, determine the value of the input voltage for which $V_{OUT} = V_{IN}$.

P5.2. For the *unloaded* RTL inverter of Figure 5.83, $V_{IN} = 1.1$ V as shown. No load gates are connected.

1. Determine the mode of operation for the transistor.
2. Determine the supply current I_{CC}.
3. Determine the output voltage V_{OUT}.

P5.3. For the RTL NOR gate of Figure 5.84, the inputs are as shown. Assume $N = 5$.

1. Determine the mode of operation for each of the three transistors.
2. Determine the supply current I_{CC}.
3. Determine the output voltage V_{OUT}.

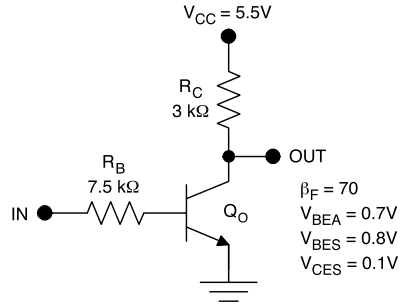


FIGURE 5.82
Unloaded RTL inverter (P5.1).

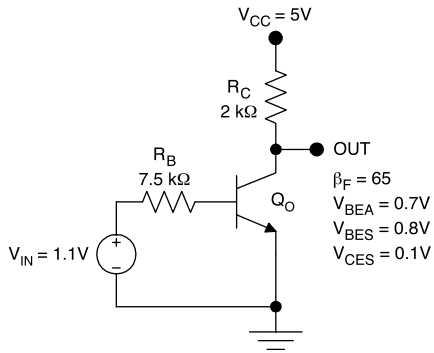


FIGURE 5.83
Unloaded RTL inverter (P5.2).

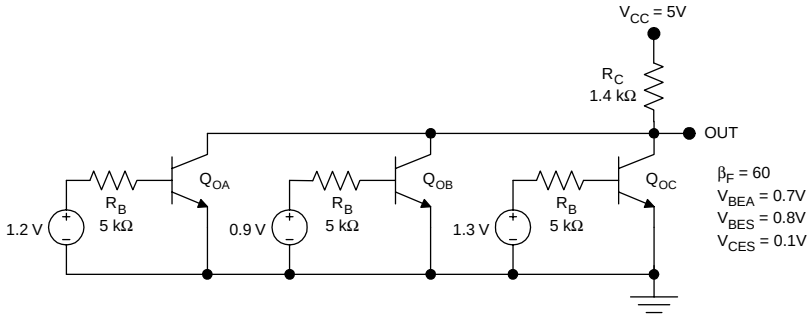


FIGURE 5.84
RTL NOR gate (P5.3).

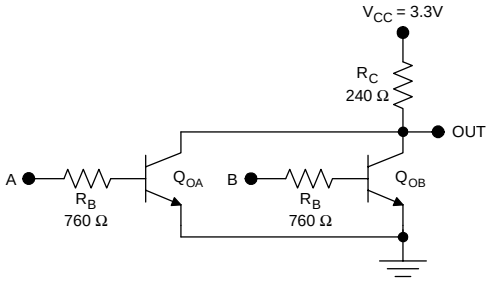


FIGURE 5.85
RTL NOR gate (P5.4).

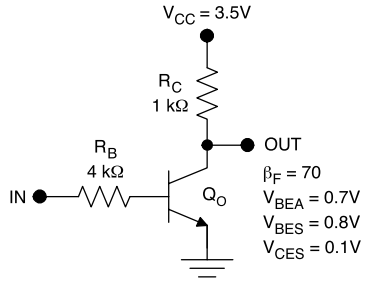


FIGURE 5.86
RTL gate (P5.5).

P5.4. For the RTL NOR gate of Figure 5.85, determine the maximum fan-out assuming that the minimum value of the high noise margin is 0.1 V. $\beta_F = 70$.

P5.5. Consider the RTL gate shown in Figure 5.86.

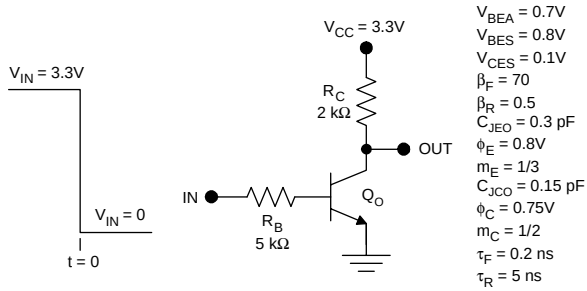


FIGURE 5.87
Unloaded RTL gate (P5.6).

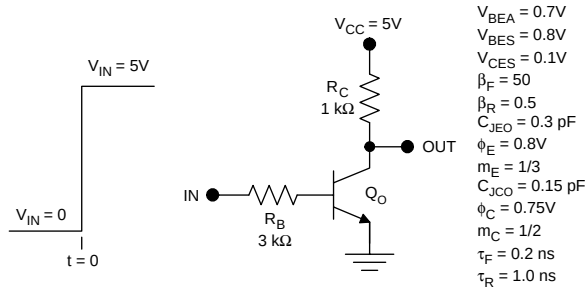


FIGURE 5.88
Unloaded RTL gate (P5.7).

1. Use SPICE to calculate and plot the unloaded VTC for this gate.
2. Hand calculate the unloaded VTC and plot it on the same graph as the SPICE results. Compare the SPICE and hand calculations.

P5.6. Consider the *unloaded* RTL gate of Figure 5.87.

1. Estimate the saturation delay.
2. Estimate the rise time.
3. Estimate the low-to-high propagation delay.

P5.7. Consider the *unloaded* RTL gate depicted in Figure 5.88.

1. Estimate the delay time.
2. Estimate the fall time.
3. Estimate the high-to-low propagation delay.

P5.8. Consider the RTL inverter of Figure 5.89 with a *lumped capacitive load*.

1. Estimate the saturation delay.
2. Estimate the low-to-high propagation delay.

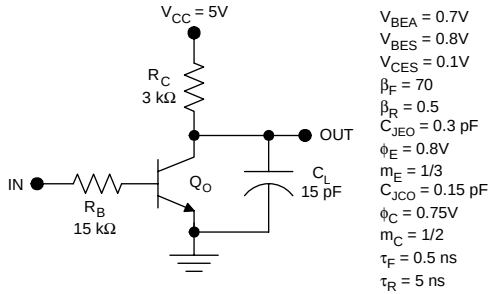


FIGURE 5.89
RTL inverter (P5.8).

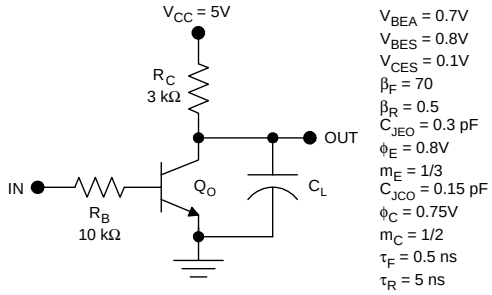


FIGURE 5.90
RTL inverter (P5.9).

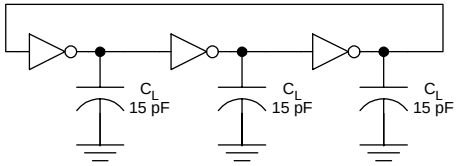


FIGURE 5.91
Three-stage ring oscillator (P5.10).

P5.9. For the RTL inverter of Figure 5.90 with a *lumped capacitive load*, determine and plot the low-to-high propagation delay as a function of C_L.

P5.10. Consider a three-stage ring oscillator, as shown in Figure 5.91, constructed using DTL inverters as shown in Figure 5.92.

1. Using hand calculations, estimate the oscillation frequency for a three-stage ring oscillator constructed using these gates. Assume that each stage is loaded by 15 pF.
2. Using SPICE, determine the oscillation frequency for the same three-stage ring oscillator.

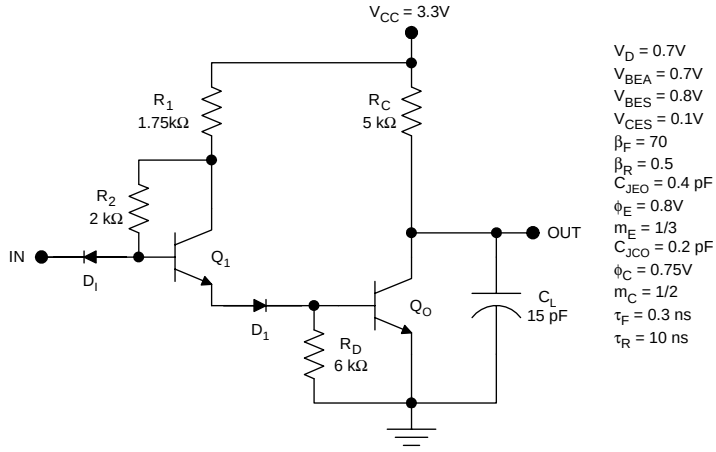


FIGURE 5.92
Three-stage ring oscillator constructed using DTL inverters (P5.10).

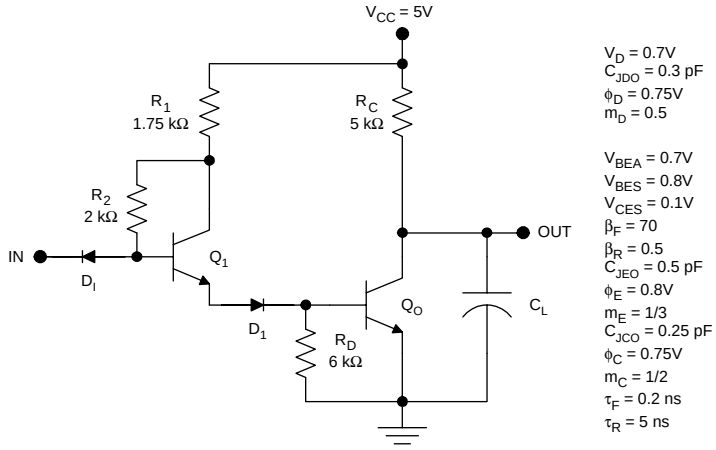


FIGURE 5.93
DTL gate (P5.11).

P5.11. Consider the DTL gate of Figure 5.93 with a *lumped capacitive load*.
1. Using SPICE, determine and plot t_{PLH} vs. C_L.
2. Using SPICE, determine and plot t_{PHL} vs. C_L.
3. Determine and plot the average propagation delay vs. C_L.

P5.12. Consider the TTL inverter illustrated in Figure 5.94.
1. Determine V_{IL}, V_{IH}, V_{OL}, and V_{OH}.
2. Determine P_L, P_H, and the average DC dissipation.
3. Determine I_{IL} and I_{IH}.

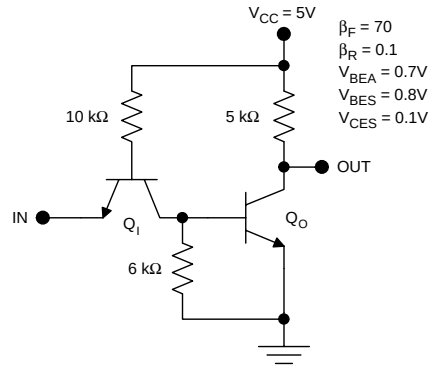


FIGURE 5.94
TTL inverter (P5.12).

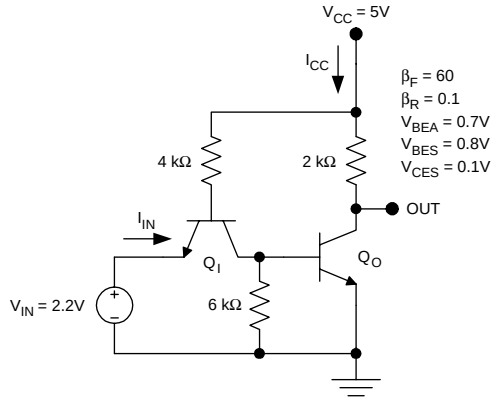


FIGURE 5.95
TTL inverter with steady voltage applied at input (P5.13).

P5.13. Consider the TTL inverter with a steady voltage applied at the input as shown in Figure 5.95.

1. Determine the mode of operation for each of the transistors.
2. Determine the supply current I_{CC}.
3. Determine the input current I_{IN}.
4. Determine the output voltage.

P5.14. Consider the standard TTL inverter of Figure 5.96.

1. Determine and plot the voltage transfer characteristic, showing all critical voltages.
2. Determine P_L, P_H, and the average DC dissipation.
3. Determine I_{IL} and I_{IH}.
4. Determine the maximum fan-out based on DC considerations, assuming $\sigma_{MAX} = 1/2$.

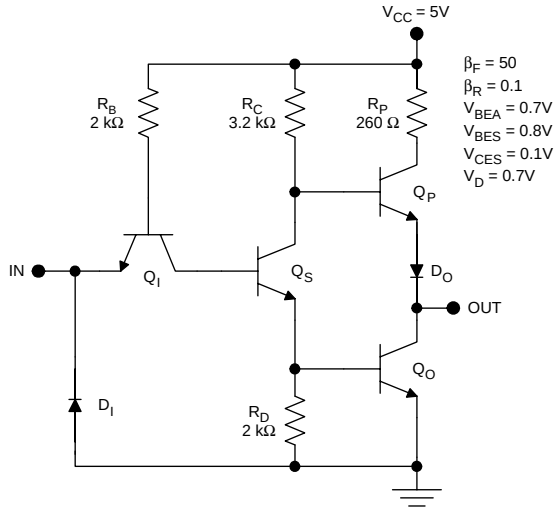


FIGURE 5.96
Standard TTL inverter (P5.14).

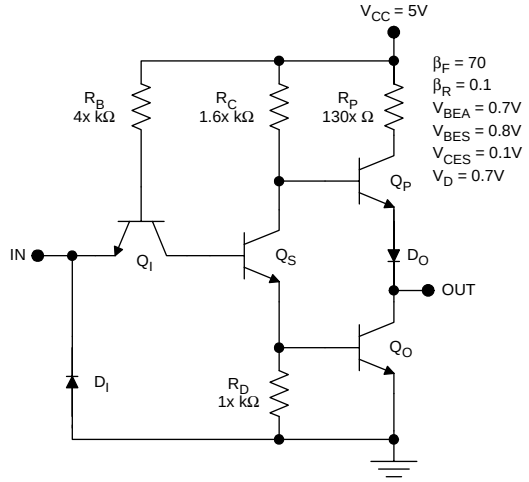


FIGURE 5.97
Standard TTL inverter with scaled resistors (P5.15).

P5.15. For the standard TTL inverter of Figure 5.97 with scaled resistors, calculate and plot I_{CCL}, I_{CCH}, I_{IL}, and I_{IH} as functions of x. $1/4 < x < 4$.

P5.16. Consider the TTL inverter of Figure 5.98 with scaled resistors. $1/4 < x < 4$.

1. Determine and plot P_H and P_L vs. x.
2. Determine and plot t_{PLH} vs. x.

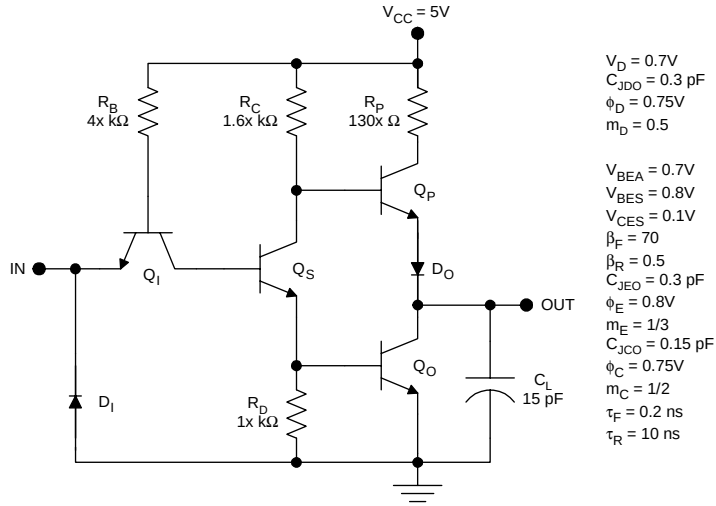


FIGURE 5.98
TTL inverter with scaled resistors (P5.16).

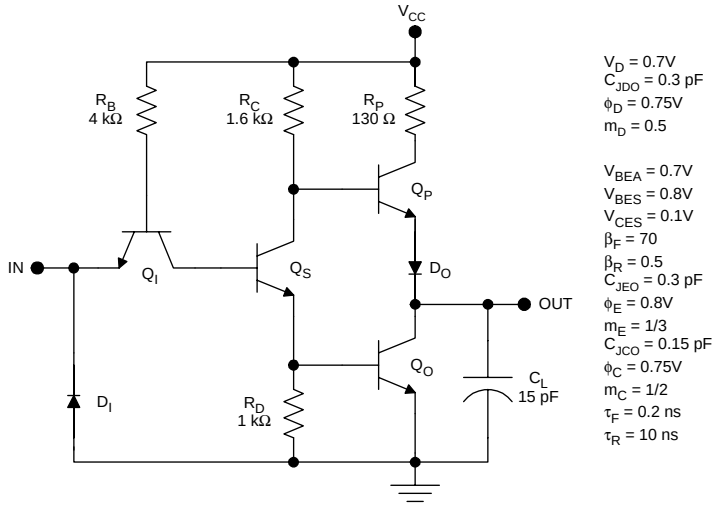


FIGURE 5.99
TTL inverter (P5.17).

3. Determine and plot t_{PHL} vs. x.
4. Determine and plot the power delay product vs. x.

P5.17. Consider the TTL inverter shown in Figure 5.99. 2.5 V < V_{CC} < 10 V.

1. Determine and plot P_H and P_L vs. V_{CC}.
2. Determine and plot t_{PLH} vs. V_{CC}.

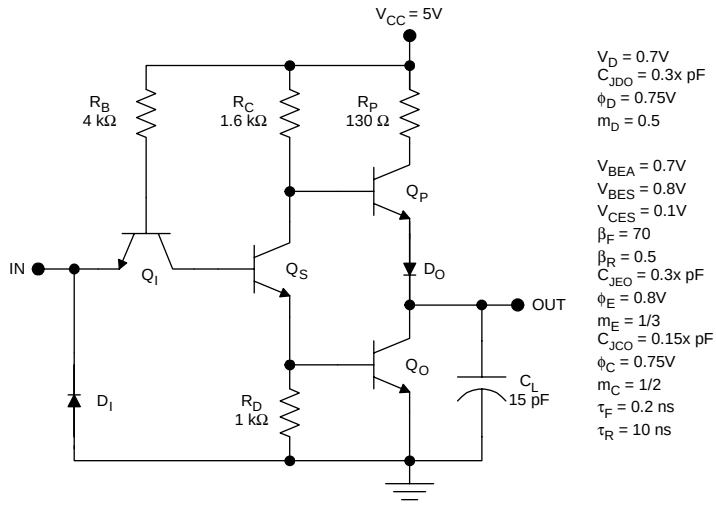


FIGURE 5.100
Standard TTL inverter (P5.18).

3. Determine and plot t_{PHL} vs. V_{CC}.
4. Determine and plot the power delay product vs. V_{CC}.

P5.18. Consider the standard TTL inverter of Figure 5.100. All transistors and diodes may be scaled in geometrical size with a direct effect on the parasitic capacitances C_{JDO}, C_{JEO}, and C_{JCO} as shown in the figure.

1. Using SPICE, determine and plot t_{PLH} vs. x.
2. Using SPICE, determine and plot t_{PHL} vs. x.
3. Determine and plot the power delay product vs. x.

P5.19. Consider the Schottky TTL gate shown in Figure 5.101.

1. Determine V_{IL}, V_{IH}, V_{OL}, and V_{OH} and plot the voltage transfer characteristic.
2. Determine I_{IL} and I_{IH}.
3. Calculate P_H, P_L, and the average DC power dissipation.

P5.20. Consider the Schottky TTL gate of Figure 5.102.

1. Determine and plot the input current I_{IN} vs. the input voltage.
2. Determine and plot the supply current I_{CC} vs. the input voltage.

P5.21. Consider the Schottky TTL inverter with a steady DC input as shown in Figure 5.103.

1. Determine the mode of operation for each of the transistors.
2. Determine the value of the input current. Use the convention that the input current is positive if it flows into the emitter of Q_I.

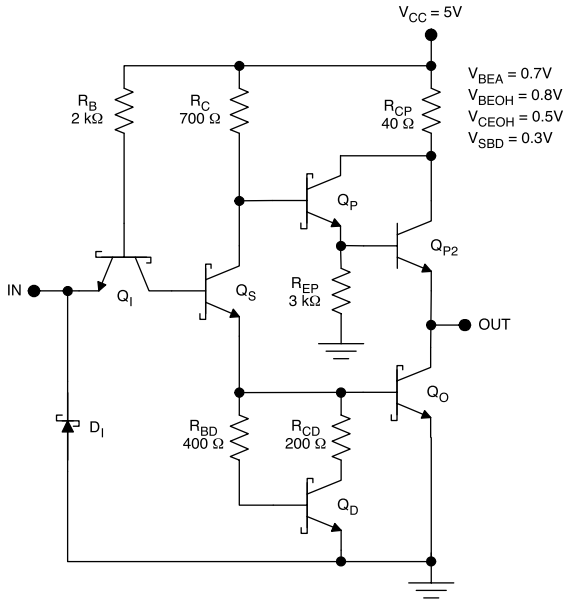


FIGURE 5.101
Schottky TTL gate (P5.19).

3. Determine the supply current I_{CC}.
4. Determine V_{OUT}.

P5.22. Consider the inverter shown in Figure 5.104.

1. Determine t_{PLH} and t_{PHL} for the circuit as shown.
2. Determine t_{PLH} and t_{PHL} assuming that the Schottky clamping has been removed from each transistor.

P5.23. Consider the Schottky TTL inverter illustrated in Figure 5.105. 5 pF < C_L < 50 pF.

1. Using SPICE, determine and plot t_{PLH} vs. C_L. Determine the average slope of the characteristic in ohms.
2. Using SPICE, determine and plot t_{PHL} vs. C_L. Determine the average slope of the characteristic in ohms.
3. Determine and plot the PDP product vs. C_L.

P5.24. Consider the Schottky TTL inverter of Figure 5.106. 4 V < V_{CC} < 6 V. 5 pF < C_L < 50 pF.

1. Determine and plot t_{PLH} vs. C_L, with V_{CC} as a parameter.
2. Determine and plot t_{PHL} vs. C_L, with V_{CC} as a parameter.

P5.25. Consider the low-power Schottky TTL gate of Figure 5.107.

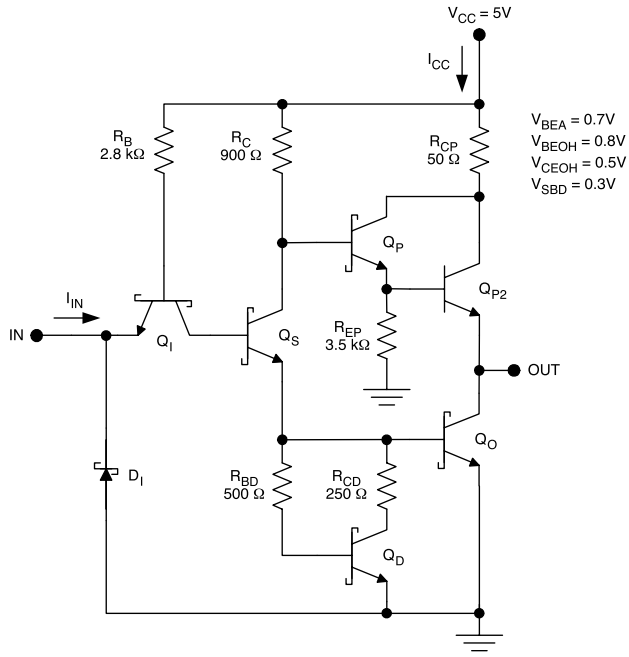


FIGURE 5.102
Schottky TTL gate (P5.20).

1. Determine V_{IL}, V_{IH}, V_{OL}, and V_{OH}.
2. Determine P_L, P_H, and the average DC dissipation.

P5.26. Consider the LSTTL gate illustrated in Figure 5.108.

1. Using SPICE, determine t_{PLH}.
2. Using SPICE, determine t_{PHL}.
3. Determine the value of the power delay product.

P5.27. Consider the low-power Schottky TTL gate of Figure 5.109 with *scaled resistors*. $1/4 < x < 4$.

1. Determine and plot P_L and P_H vs. x.
2. Determine and plot t_{PLH} and t_{PHL} vs. x.
3. Determine and plot the power delay product vs. x.

P5.28. Consider the low-power Schottky TTL circuit depicted in Figure 5.110. 5 pF $< C_L <$ 50 pF.

1. Determine and plot t_{PLH} vs. C_L. Determine the average slope of the characteristic in ohms.
2. Determine and plot t_{PHL} vs. C_L. Determine the average slope of the characteristic in ohms.

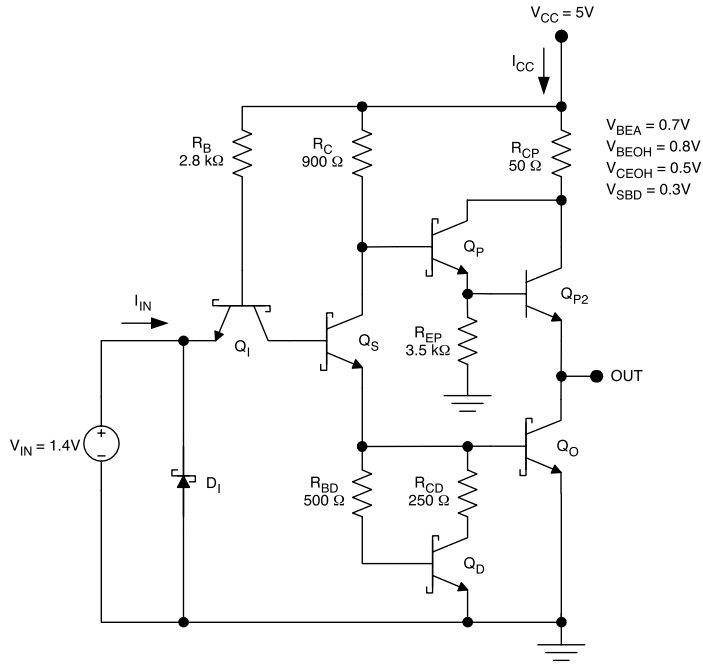


FIGURE 5.103
Schottky TTL inverter with steady DC input (P5.21).

P5.29. Consider the low-power Schottky gate of Figure 5.111 with *scaled resistors*. ¼ < x < 4.5 pF < C_L < 50 pF.

1. Determine and plot t_{PLH} vs. C_L with x as a parameter.
2. Determine and plot t_{PHL} vs. C_L with x as a parameter.

P5.30. Consider the advanced low-power Schottky TTL gate in Figure 5.112.

1. Determine V_{IL}, V_{IH}, V_{OL}, and V_{OH}.
2. Determine P_L, P_H, and the average DC dissipation.

P5.31. Consider the advanced low-power Schottky gate of Figure 5.113.

1. Using SPICE, determine t_{PLH}.
2. Using SPICE, determine t_{PHL}.
3. Determine the value of the power delay product.

P5.32. Consider the advanced low-power Schottky TTL circuit illustrated in Figure 5.114. 5 pF < C_L < 50 pF.

1. Determine and plot t_{PLH} vs. C_L. Determine the average slope of the characteristic in ohms.
2. Determine and plot t_{PHL} vs. C_L. Determine the average slope of the characteristic in ohms.

Problems continued on page 204

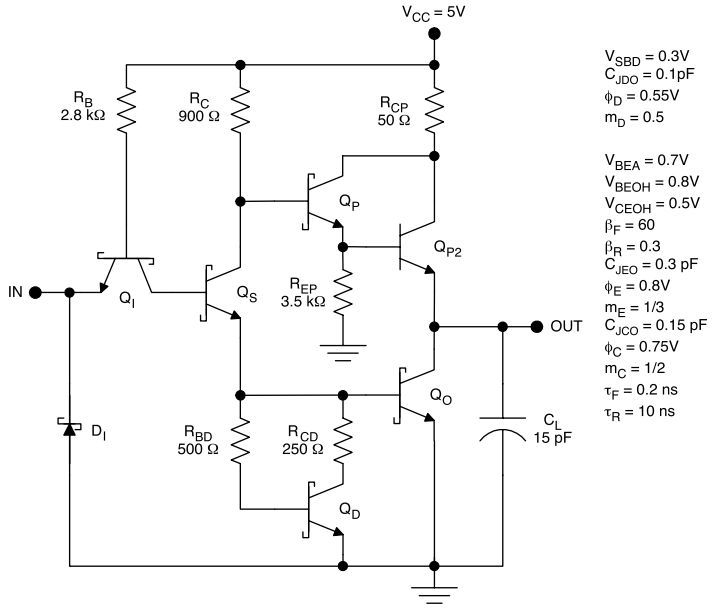


FIGURE 5.104
Inverter (P5.22).

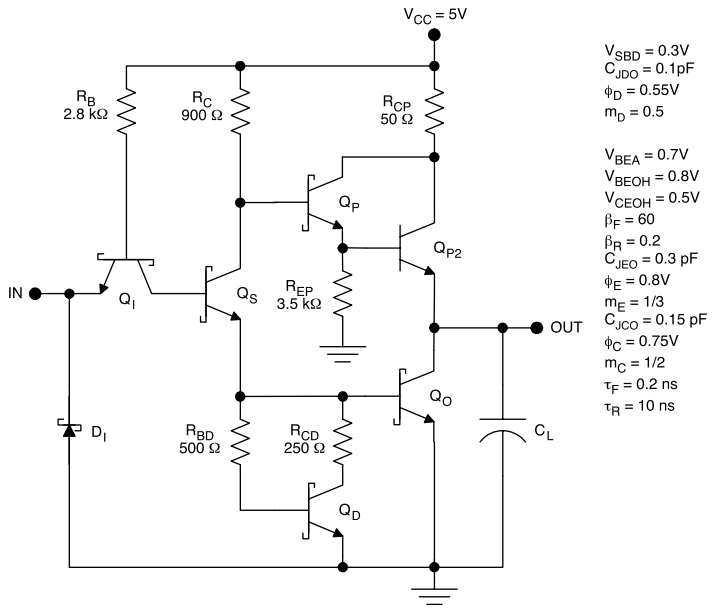


FIGURE 5.105
Schottky TTL inverter (P5.23).

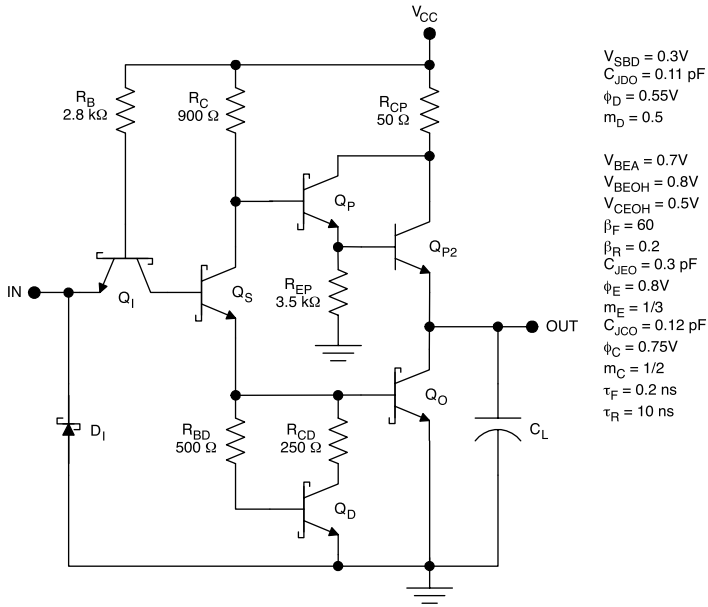


FIGURE 5.106
Schottky TTL inverter (P5.24).

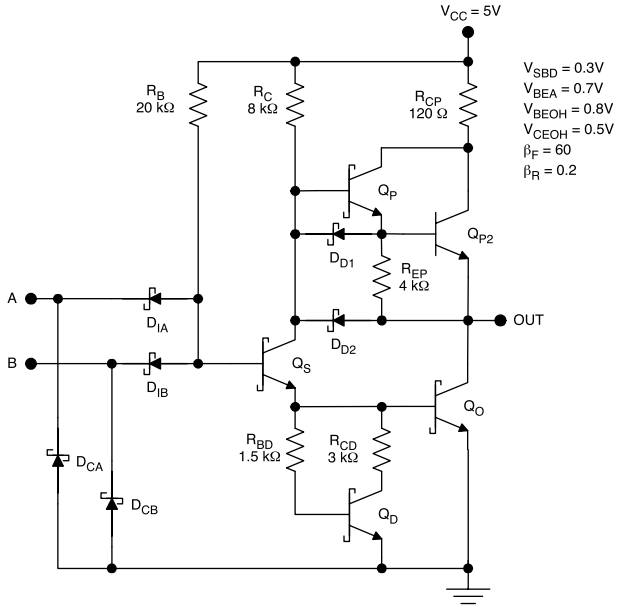


FIGURE 5.107
Low-power Schottky TTL gate (P5.25).

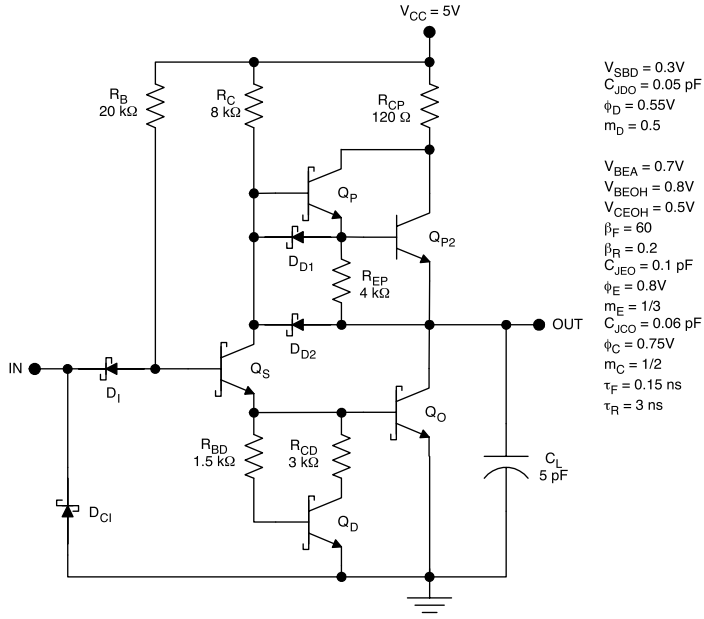


FIGURE 5.108
LSTTL gate (P5.26).

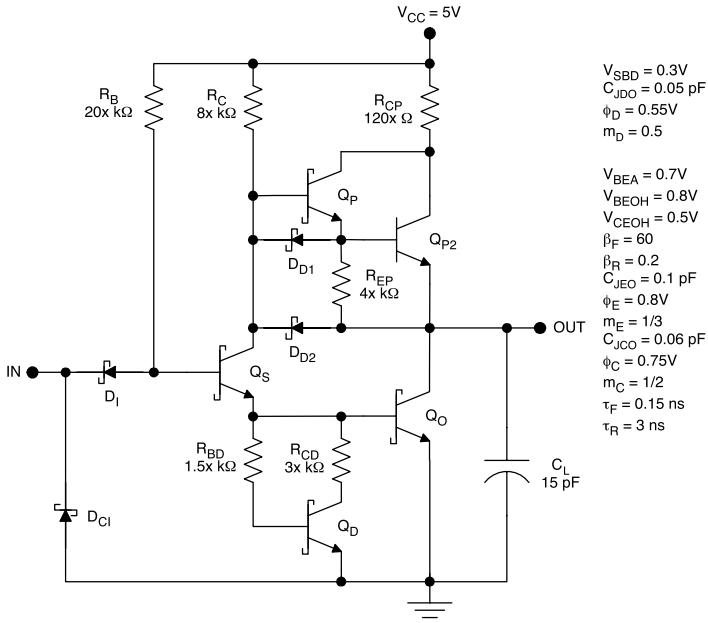


FIGURE 5.109
Low-power Schottky TTL gate (P5.27).

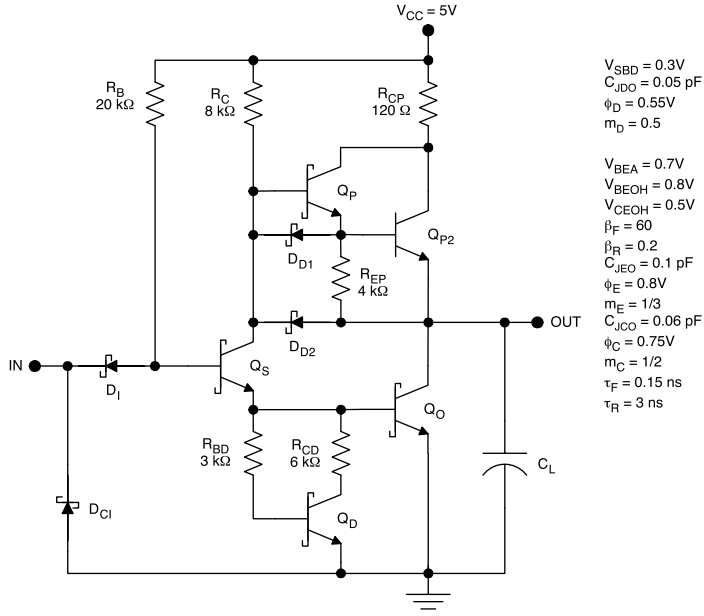


FIGURE 5.110
Low-power Schottky TTL circuit (P5.28).

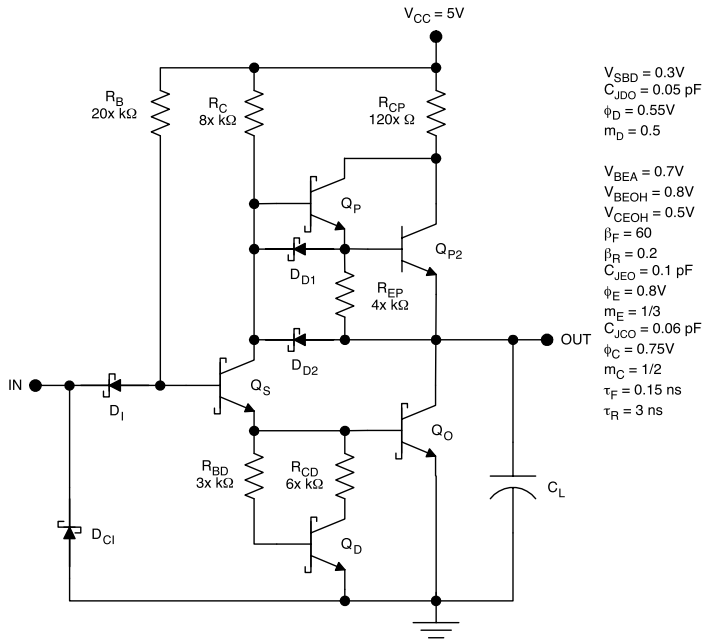


FIGURE 5.111
Low-power Schottky gate (P5.29).

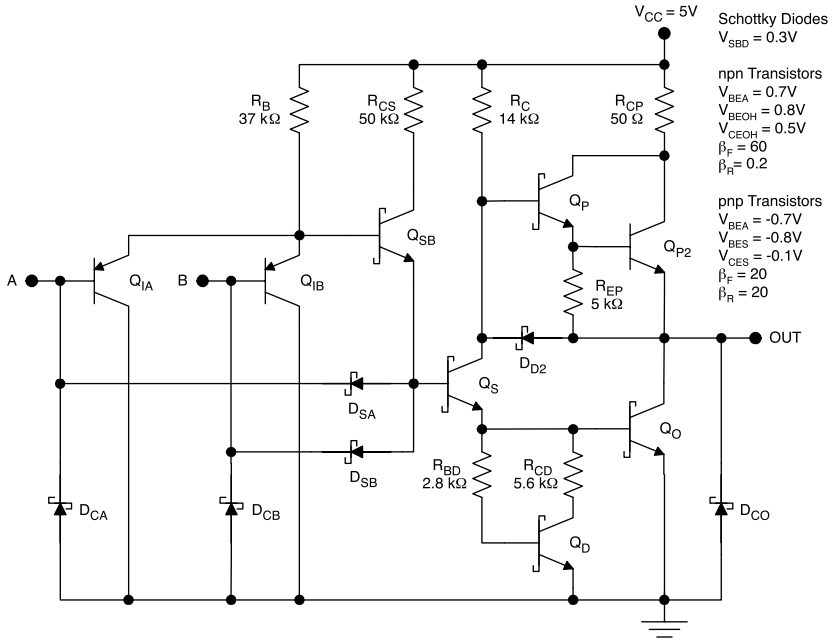


FIGURE 5.112
Advanced low-power Schottky TTL gate (P5.30).

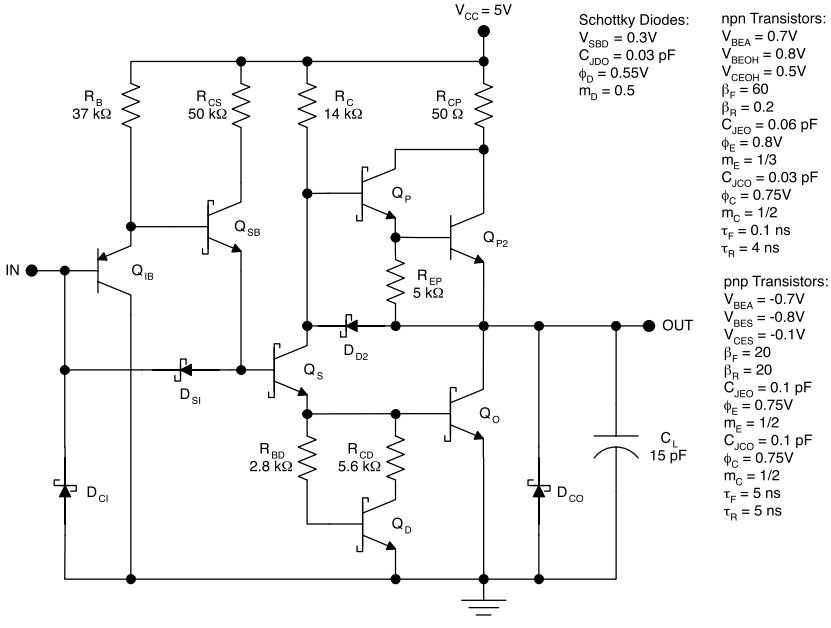


FIGURE 5.113
Advanced low-power Schottky gate (P5.31).

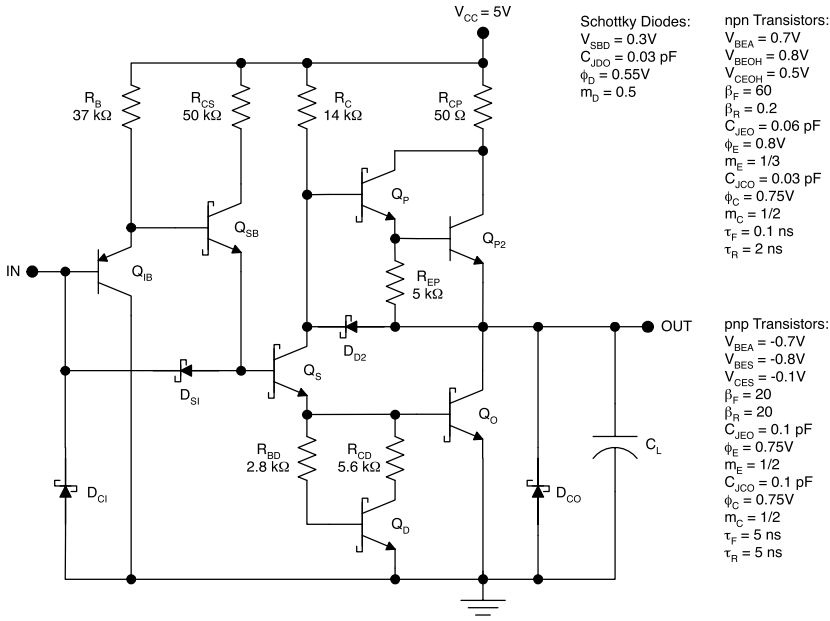


FIGURE 5.114
Advanced low-power Schottky TTL circuit (P5.32).

P5.33. Consider the Fairchild advanced Schottky TTL gate of Figure 5.115
1. Determine V_{IL}, V_{IH}, V_{OL}, and V_{OH}.
2. Determine P_L, P_H, and the average DC dissipation.

P5.34. Consider the Fairchild advanced Schottky TTL gate illustrated in Figure 5.116.
1. Using SPICE, determine t_{PLH}.
2. Using SPICE, determine t_{PHL}.
3. Determine the value of the power delay product.

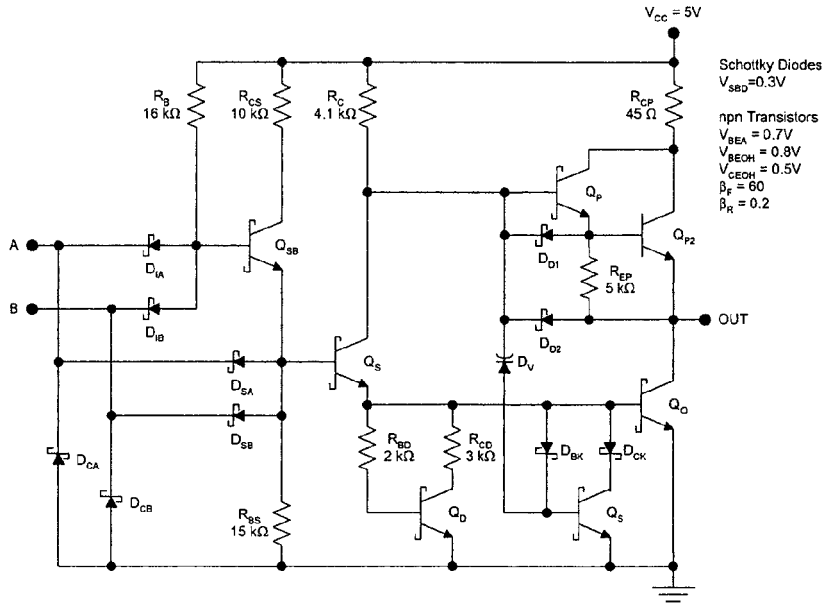


FIGURE 5.115
Fairchild advanced Schottky TTL gate (P5.33).

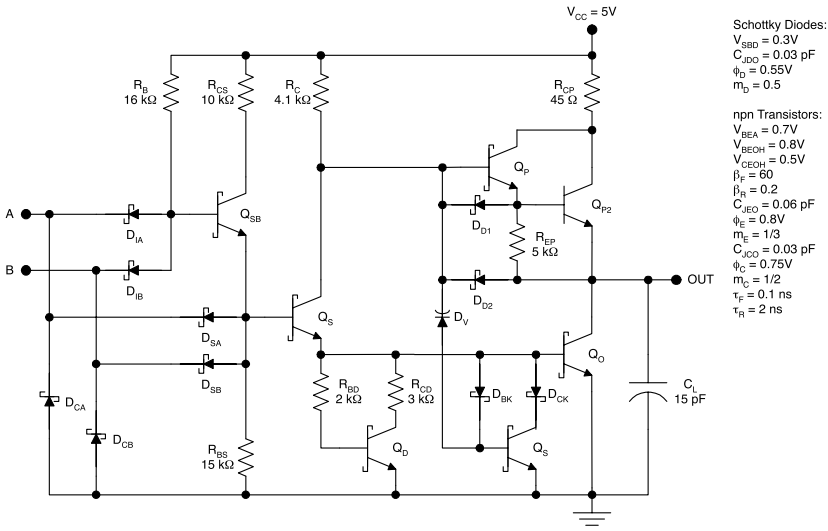


FIGURE 5.116
Fairchild advanced Schottky TTL gate (P5.34).

References

1. www.ti.com (Texas Instruments).
2. www.fairchildsemi.com (Fairchild Semiconductor).
3. www.cadence.com (Cadence).

6

Emitter-Coupled Logic

6.1 Introduction

Emitter-coupled logic (ECL) circuits are the fastest available bipolar logic gates. At the present time, standard ECL circuits achieve propagation delays of less than 25 ps.[1,2] Compared to other logic families, these circuits are less susceptible to capacitive loading because of their extremely low output impedance. Therefore, they can drive off-chip loads such as busses or transmission lines at data rates greater than 10 GHz. These advantages have been exploited extensively in the high-speed digital communication and supercomputing applications that necessitate the raw speed of ECL.

The speed of ECL stems from several unique aspects of the circuit design. First, although the other bipolar logic families are based on voltage mode logic, ECL utilizes *current mode logic*. Here the important distinction is that TTL uses bipolar transistors as voltage mode switches that swing from cutoff to saturation. In contrast, ECL utilizes bipolar transistors as current mode switches that never saturate. This approach avoids the capacitive effects and saturation delays associated with excess minority carrier storage in the devices. A second advantage of ECL is that small logic swings are used (less than 1 V), resulting in short rise and fall times. Finally, the output drivers used in ECL are emitter followers that are always on and achieve minimal output impedance, therefore providing superior immunity to capacitive loading at the output.

The primary disadvantage of ECL is the high quiescent power dissipation. The current mode logic concept requires that a steady DC current flow in an emitter-coupled current switch (described in the next section). The emitter follower output drivers also require large and steady emitter currents. This stems from the fact that the output impedance of an emitter follower is approximately

$$R_{\text{OUT}} \approx \frac{1}{g_m} = \frac{kT}{qI_E}, \tag{6.1}$$

where g_m is the transconductance of the bipolar transistor, k is the Boltzmann constant, T is the temperature, q is the electronic charge, and I_E is the DC emitter current. Thus a direct trade-off exists between the levels of current in the emitter followers and the output impedance, which translates into the usual speed–power trade-off. The ECL gate draws a significant current flow in either logic state and dissipates a considerable amount of DC power — usually in the tens of milliwatts. This seemingly small amount of dissipation would be problematic in a processor comprising 1 million gates or more. Therefore, a disadvantage of ECL is that considerations of heat removal place a practical limit on the number of gates that can be incorporated in a single chip of silicon.

Historically, ECL had been the mainstay of high-end supercomputers; however, the high dissipation required the implementation of elaborate cooling schemes for ECL processor cores. Typically, forced water cooling was implemented using microchannels cut in the silicon wafers. The drawbacks of this approach include the need for complicated packaging, bulky cooling equipment, and the associated maintenance requirements.

The ever improving speed of CMOS circuitry coupled with massively parallel computing architectures has made it possible to implement supercomputer processing cores in CMOS or BiCMOS. This approach has greatly reduced the processor dissipation and associated cooling requirements, resulting in smaller and more cost-effective machines.

Nonetheless, the move to CMOS in parallel processor machines has not obviated the need for ECL. Bipolar gates such as ECL achieve superior data rates in high-capacity interchip and intermodule links *within* massively parallel machines. They are also important in other critical digital communication applications. Processors and gate arrays continue to be implemented in ECL for critical applications requiring high speed and high off-chip bitrates.[3–7] At the present time, heterojunction bipolar transistors (HBTs) based on $Si_{1-x}Ge_x$ are poised to improve the performance of ECL gates and the achievable data rates further[8–11]

6.2 Circuit Evolution

The key subcircuit in any ECL gate is the emitter-coupled current switch, the simplest form of which is shown in Figure 6.1. In this simple current switch, V_{REF} is a constant reference voltage. A nearly constant current I_{EE} flows in the resistor R_E. However, this current is diverted entirely through one of the two emitter-coupled transistors, depending on the value of the input voltage. If the input voltage is slightly greater than the reference voltage, essentially all of the current I_{EE} flows through Q_1, which is forward active. Under this condition of a logic-one input, the output voltage swings low to

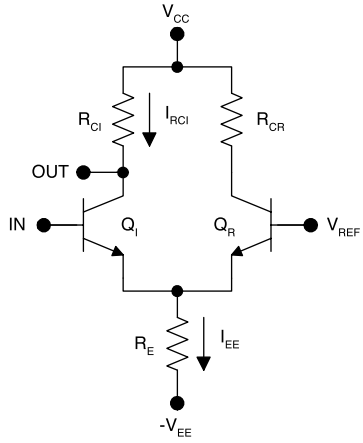


FIGURE 6.1
Emitter-coupled current switch.

$$V_{\text{OUT}} = V_{CC} - I_{RCI}R_{CI} \approx V_{CC} - I_{EE}R_{CI}\,. \tag{6.2}$$

On the other hand, if the input voltage is slightly less than the reference voltage, essentially all of the current flows through R_{CR}. Under this condition of logic-zero input, the output goes high to

$$V_{\text{OUT}} = V_{CC} - I_{RCI}R_{CI} \approx V_{CC}\,. \tag{6.3}$$

Therefore, the emitter-coupled current switch is an inverter with a logic swing of

$$V_{OH} - V_{OL} = I_{EE}R_{CI}\,, \tag{6.4}$$

the value of which can be established by the proper choices of the resistor values in the circuit.

In the aforementioned current switch, the output voltage levels are referenced to V_{CC}. The most stable value of V_{CC} is ground, as shown in the revised circuit of Figure 6.2. Another advantage of using ground for the collector supply is that this level is more consistent in large systems that have independently powered subsystems. An important consequence of this approach is that all of the voltage levels are negative.

A basic ECL gate can be constructed using the emitter-coupled current switch, as shown in Figure 6.3. Two important aspects of this circuit are the inclusion of emitter followers and the provision for complementary outputs. The emitter followers Q_{INV} and Q_{NINV} provide very low output impedance (a few ohms) and diminish the effect of capacitive loading; unlike the emitter followers used in TTL circuitry, these transistors are always conducting. Therefore, they provide this benefit over the entire range of output. A secondary

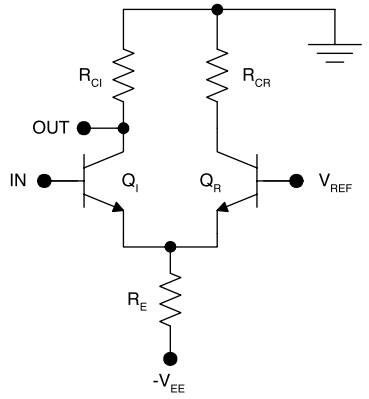


FIGURE 6.2
Emitter-coupled current switch with output levels referenced to ground.

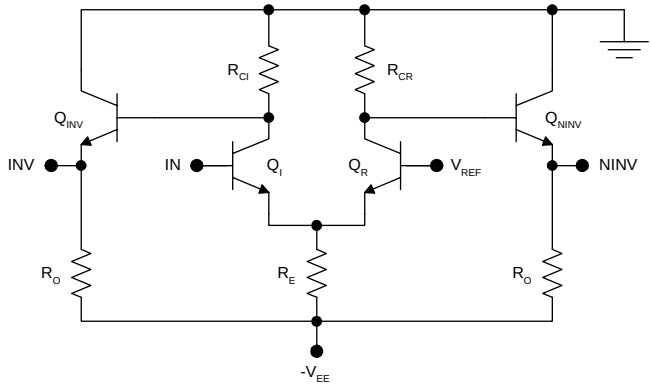


FIGURE 6.3
Emitter-coupled logic gate with complementary outputs.

benefit of the emitter followers is that they decrease both DC output levels by V_{BEA}. This is helpful in keeping the current switch transistors from saturating when the output levels are applied to similar gates. Complementary outputs are always provided in ECL gates because they add extra functionality at little cost. Thus, V_{INV} is the inverting output and V_{NINV} is the noninverting output.

All ECL circuitry in use today is essentially similar to the gate shown in Figure 6.3. Important circuit modifications make the modern gates more immune to variations in temperature or supply voltage. These include replacement of R_E with a current source, addition of temperature compensation diodes to R_{CI} and R_{CR}, and use of a temperature-compensated bias driver instead of a constant reference voltage. Nonetheless, these improvements keep the core of the circuit, including the emitter-coupled current switch and the emitter follower output drivers, intact.

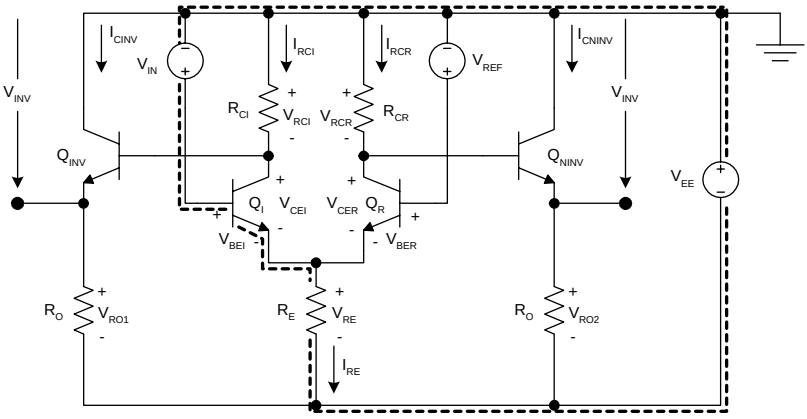


FIGURE 6.4
ECL circuit for determination of I_{RE} using KVL.

6.3 Using Kirchhoff's Voltage Law with ECL Circuits

The currents and voltages in an ECL circuit may be determined using Kirchhoff's voltage law (KVL) and Ohm's law. Consider the ECL circuit of Figure 6.4 and suppose that logic one is applied at the input so that Q_I is forward active. To determine the current in the emitter resistor, KVL can be used for the loop indicated. Proceeding around the loop in a counterclockwise direction, a voltage term is added if the polarity is such that the plus sign is encountered first; otherwise the term is subtracted. For the loop indicated, this yields

$$-V_{RE} - V_{BEI} + V_{IN} + V_{EE} = 0\,. \tag{6.5}$$

Using Ohm's law for R_E yields

$$-I_{RE}R_E - V_{BEI} + V_{IN} + V_{EE} = 0\,. \tag{6.6}$$

Rearranging yields an expression for the desired current:

$$I_{RE} = \frac{V_{EE} + V_{IN} - V_{BEI}}{R_E} = \frac{V_{EE} + V_{IN} - V_{BEA}}{R_E}\,. \tag{6.7}$$

Normally, the first two preceding steps are skipped. Also, simplified circuit diagrams are used in which the loops for the KVL analysis may not always be shown explicitly. Following this approach with the simplified circuit and the path highlighted, in Figure 6.5, the desired current is obtained in one step:

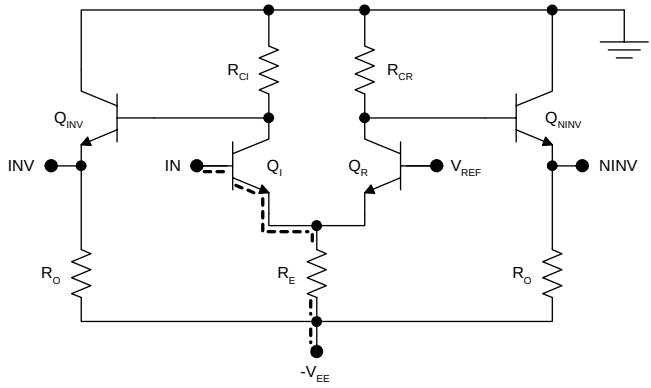


FIGURE 6.5
Simplified ECL for use of KVL shortcut method to determine current in resistor R_E.

$$I_{RE} = \frac{V_{IN} + V_{EE} - V_{BEA}}{R_E}. \tag{6.8}$$

Because this KVL shortcut approach is more direct and avoids the use of cumbersome circuit diagrams, it is used throughout this book.

6.4 Voltage Transfer Characteristic

Consider the noninverting output for the gate shown in Figure 6.6. If the input is well below the reference voltage, then Q_I is cut off and Q_R is forward active. The emitter current flowing in Q_R is

$$I_{RE} = \frac{V_{REF} - V_{BEA} - \left(-V_{EE}\right)}{R_E}. \tag{6.9}$$

The collector current flowing in Q_R is

$$I_{RCR} = \left(\frac{\beta_F}{1+\beta_F}\right)\frac{V_{REF} - V_{BEA} - \left(-V_{EE}\right)}{R_E}, \tag{6.10}$$

so the voltage at the base of the output transistor Q_{NINV} is

$$V_{BNINV} = 0 - R_{CR}\left(\frac{\beta_F}{1+\beta_F}\right)\frac{V_{REF} - V_{BEA} - \left(-V_{EE}\right)}{R_E}. \tag{6.11}$$

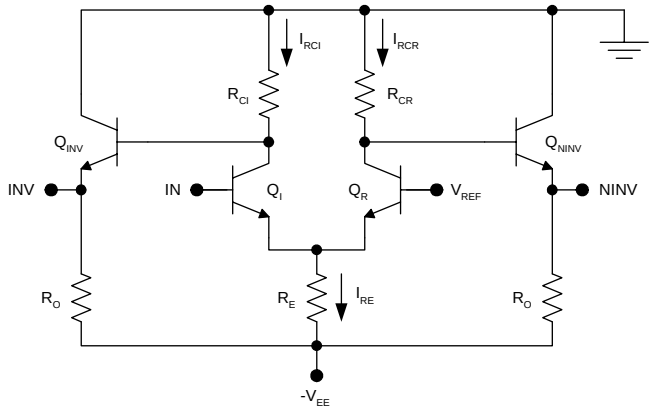


FIGURE 6.6
ECL gate for determination of voltage transfer characteristics.

The low output voltage at the noninverting output is therefore

$$V_{OL} = -R_{CR}\left(\frac{\beta_F}{1+\beta_F}\right)\frac{V_{REF} - V_{BEA} - \left(-V_{EE}\right)}{R_E} - V_{BEA}\,. \tag{6.12}$$

Typically, the resistor values are chosen, so this value is approximately $-2\,V_{BEA}$.

Q_I will begin to conduct when the input voltage approaches V_{REF} because its base–emitter voltage will approach V_{BEA}. Further increase of the input voltage above V_{REF} will cause Q_R to cut off so that all the current will be carried by Q_I. The width of this transition is very narrow because the emitter current in a bipolar transistor increases by a decade for each 60 mV increase in the base–emitter voltage. It is customary to assume that the width of the transition is 1/10 of a volt. Then,

$$V_{IL} \approx V_{REF} - 0.05\ \text{V} \tag{6.13}$$

and

$$V_{IH} \approx V_{REF} + 0.05\ \text{V}\,. \tag{6.14}$$

With the input greater than V_{IH}, the reference transistor Q_R is cut off; therefore, if the base current for the emitter follower transistor is neglected,

$$V_{OH} = -V_{BEA}\,. \tag{6.15}$$

Now consider the voltage transfer characteristic for the inverting output. The critical input voltages are the same as for the noninverting output as is

the output high level. There are some important differences for the case of a low output, however. If the input transistor is conducting, then the emitter current is

$$I_{RE} = \frac{V_{IN} - V_{BEA} - \left(-V_{EE}\right)}{R_E}. \tag{6.16}$$

If the input transistor is forward active, then its collector current is

$$I_{RCI} = \left(\frac{\beta_F}{1+\beta_F}\right)\frac{V_{IN} - V_{BEA} - \left(-V_{EE}\right)}{R_E} \tag{6.17}$$

and the inverting output voltage is

$$V_{INV} = 0 - R_{CI}\left(\frac{\beta_F}{1+\beta_F}\right)\frac{V_{IN} - V_{BEA} - \left(-V_{EE}\right)}{R_E} - V_{BEA}. \tag{6.18}$$

The output low voltage is the value corresponding to the application of the output high voltage from a similar gate. Thus,

$$V_{OL} = -R_{CI}\left(\frac{\beta_F}{1+\beta_F}\right)\frac{V_{OH} - V_{BEA} - \left(-V_{EE}\right)}{R_E} - V_{BEA}. \tag{6.19}$$

If the input voltage is sufficiently high, the input transistor will saturate, creating an anomaly in the transfer characteristic. This occurs when the collector–emitter voltage for Q_I decreases to the saturated level V_{CES}. With Q_I forward active, the voltage at the emitter is

$$V_{EI} = V_{IN} - V_{BEA} \tag{6.20}$$

and the voltage at the collector of Q_I is

$$V_{CI} = -R_{CI}\left(\frac{\beta_F}{1+\beta_F}\right)\frac{V_{IN} - V_{BEA} - \left(-V_{EE}\right)}{R_E}. \tag{6.21}$$

Therefore, the onset of saturation in Q_I corresponds to the condition

$$V_{CE} = -R_{CI}\left(\frac{\beta_F}{1+\beta_F}\right)\frac{V_{IN} - V_{BEA} - \left(-V_{EE}\right)}{R_E} - V_{IN} + V_{BEA} = V_{CES}. \tag{6.22}$$

Solving the equation determines that the input transistor saturates with

$$V_{IN} = \frac{\left(\frac{\beta_F}{1+\beta_F}\right)\frac{R_{CI}}{R_E}\left(V_{BEA} - V_{EE}\right) - V_{CES} + V_{BEA}}{1+\left(\frac{\beta_F}{1+\beta_F}\right)\frac{R_{CI}}{R_E}}. \tag{6.23}$$

Once the input transistor saturates, its base–collector junction becomes forward biased. Under this condition the inverting transfer characteristic is given by

$$V_{\text{OUT}} = V_{IN} - V_{BCS} - V_{BEA}. \tag{6.24}$$

Thus the output voltage increases linearly with V_{IN} under this (undesirable) condition of saturation in the input transistor.

Example 6.1

Calculate the critical voltages for the ECL gate of Figure 6.7.

Solution. Using the 1/10 V rule of thumb, the critical input voltages are $V_{IL} \approx V_{REF} - 0.05$ V = –1.225 V and $V_{IH} \approx V_{REF} + 0.05$ V = –1.125 V.

Now consider the noninverting output levels. With Q_R cut off, a negligible voltage drop in R_{CR} occurs so that $V_{OH} = -V_{BEA} = -0.75$ V. With Q_R conducting, the emitter current is

$$I_{RE} = \frac{V_{REF} - V_{BEA} - \left(-V_{EE}\right)}{R_E} = \frac{-1.175\text{ V} - 0.75\text{ V} - \left(-5.2\text{ V}\right)}{1.24\text{ k}\Omega} = 2.64\text{ mA}$$

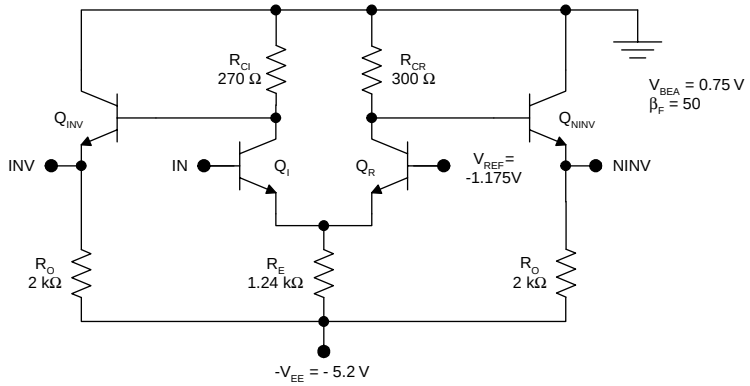


FIGURE 6.7
Example of ECL gate for calculation of critical voltages.

TABLE 6.1

Critical Voltages for the ECL Gate Example

Symbol	Noninverting Output	Inverting Output
V_{IL}	–1.225 V	–1.225 V
V_{IH}	–1.125 V	–1.125 V
V_{OL}	–1.527 V	–1.538 V
V_{OH}	–0.75 V	–0.75 V

and the collector current is

$$I_{RCR} = \left(\frac{\beta_F}{1+\beta_F}\right) I_{RE} = \left(\frac{50}{51}\right) 2.64 \text{ mA} = 2.59 \text{ mA};$$

therefore, the low output level is $V_{OL} = -I_{RCR}R_{CR} - V_{BEA} = -(2.59 \text{ mA})(300\ \Omega) - 0.75 \text{ V} = -1.527 \text{ V}$.

Now consider the inverting output. With Q_I cut off, the only current flowing through R_{CI} is the base current for Q_{INV}. To the extent that this current can be neglected, the output high level is the same as for the noninverting output: $V_{OH} = -V_{BEA} = -0.75$ V. With Q_I conducting, the emitter current is

$$I_{RE} = \frac{V_{OH} - V_{BEA} - (-V_{EE})}{R_E} = \frac{-0.75 \text{ V} - 0.75 \text{ V} - (-5.2 \text{ V})}{1.24 \text{ k}\Omega} = 2.98 \text{ mA},$$

the collector current is

$$I_{RCI} = \left(\frac{\beta_F}{1+\beta_F}\right) I_{RE} = \left(\frac{50}{51}\right) 2.98 \text{ mA} = 2.92 \text{ mA},$$

and the low output level is therefore $V_{OL} = -I_{RCI}R_{CI} - V_{BEA} = -(2.92 \text{ mA})(270\ \Omega) - 0.75 \text{ V} = -1.538 \text{ V}$.

The critical voltages are summarized in Table 6.1. Notice that the circuit is designed with different values for the two collector resistors so that the output levels are nearly equal in the low output states.

Example 6.2

Determine and plot the voltage transfer characteristics for the ECL gate illustrated in Figure 6.8.

Solution. The critical input voltages are approximately $V_{IL} \approx V_{REF} - 0.05 \text{ V} = -1.225 \text{ V}$ and $V_{IH} \approx V_{REF} + 0.05 \text{ V} = -1.125 \text{ V}$. For the noninverting output, the output high voltage is $V_{OH} = -V_{BEA} = -0.75$ V and the output low voltage is

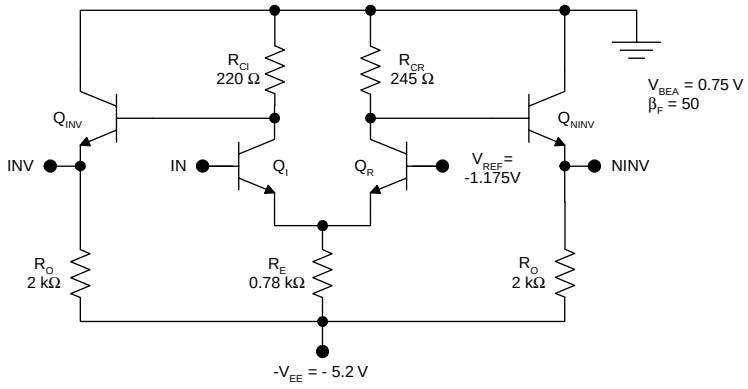


FIGURE 6.8
Example of ECL gate for determination of voltage transfer characteristics.

$$V_{OL} = -\left(\frac{\beta_F}{1+\beta_F}\right)R_{CR}\left(\frac{V_{REF} - V_{BEA} - \left(-V_{EE}\right)}{R_E}\right) - V_{BEA}$$

$$= -\left(\frac{50}{51}\right)(245\Omega)\left(\frac{-1.175\text{ V} - 0.75\text{ V} + 5.2\text{ V}}{780\Omega}\right) - 0.75\text{ V} = -1.758\text{ V}$$

The approximate noninverting transfer characteristic comprises three line segments:

$$V_{NINV} = \begin{cases} -1.758\text{ V}; & V_{IN} < -1.225\text{ V} \\ -1.758\text{ V} + 10.1\left(V_{IN} - (-1.225\text{ V})\right); & -1.225\text{ V} \le V_{IN} \le -1.125\text{ V} \\ -0.75\text{ V}; & V_{IN} > -1.125\text{ V} \end{cases}$$

The inverting transfer characteristic is more complex. For $V_{\text{IN}} < V_{IL}$, Q_I is cut off, so $V_{OH} = -V_{BEA} = -0.75$ V. The output decreases in approximately linear fashion in the transition region, $V_{IL} < V_{\text{IN}} < V_{IH}$. When the input voltage is greater than V_{IH} but not sufficient to cause the input transistor to saturate, the characteristic is given by

$$V_{INV} = -R_{CI}\left(\frac{\beta_F}{1+\beta_F}\right)\frac{V_{\text{IN}} - V_{BEA} - \left(-V_{EE}\right)}{R_E} - V_{BEA}$$

$$= -220\Omega\left(\frac{50}{51}\right)\left(\frac{V_{\text{IN}} - 0.75\text{ V} + 5.2\text{ V}}{780\Omega}\right) - 0.75\text{ V} = -1.980\text{ V} - 0.276\ V_{\text{IN}}$$

If it is assumed that the saturated collector–emitter voltage $V_{CES} = 0.05$ V, then the input transistor saturates at an input voltage given by

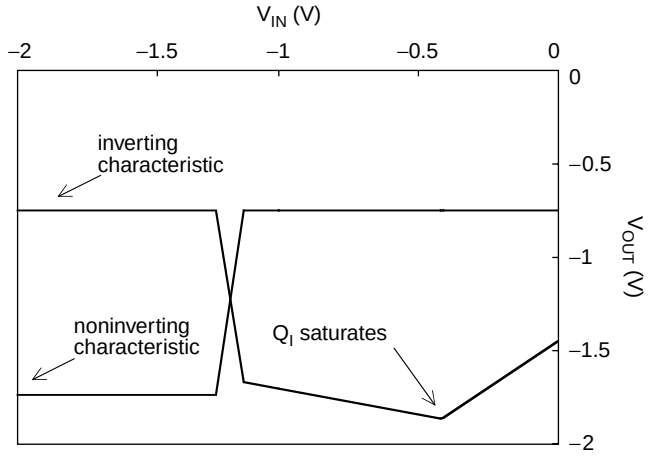


FIGURE 6.9
ECL voltage transfer characteristics.

$$V_{IN} = \frac{\left(\frac{\beta_F}{1+\beta_F}\right)\frac{R_{CI}}{R_E}\left(V_{BEA} - V_{EE}\right) - V_{CES} + V_{BEA}}{1+\left(\frac{\beta_F}{1+\beta_F}\right)\frac{R_{CI}}{R_E}}$$

$$= \frac{\left(\frac{50}{51}\right)\frac{220\Omega}{780\Omega}(0.75\text{ V} - 5.2\text{ V}) - 0.05\text{ V} + 0.75\text{ V}}{1+\left(\frac{50}{51}\right)\frac{220\Omega}{780\Omega}} = -0.42\text{ V}$$

Once the input transistor saturates, the base–collector junction becomes forward biased. If the saturated base–collector voltage V_{BCS} = 0.7 V, then, with the input transistor saturated, the inverting transfer characteristic is given by $V_{OUT} = V_{IN} - V_{BCS} - V_{BEA} = V_{IN} - 0.7\text{ V} - 0.75\text{ V} = V_{IN} - 1.45\text{ V}$.

Both voltage transfer characteristics are shown in Figure 6.9; the saturation of Q_I appears at the right side of the figure. This situation is undesirable but will not occur if the input signal is from a similar ECL gate and therefore never more positive than –0.75 V.

6.5 Dissipation

For ECL gates the DC dissipation is always dominant. The average DC dissipation is determined in the usual manner, by averaging the values for the two logic states.

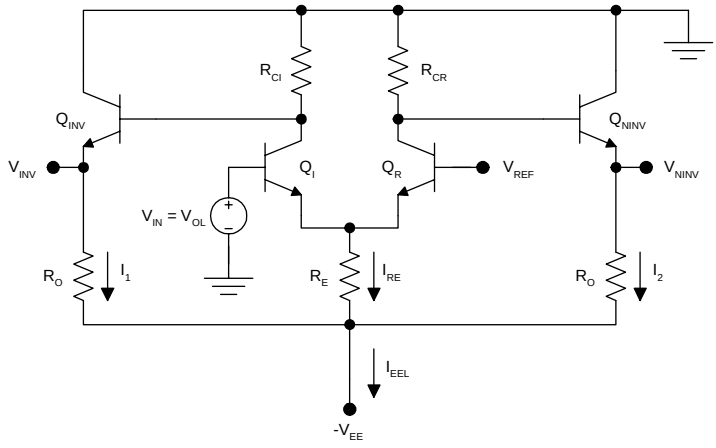


FIGURE 6.10
ECL gate for determination of P_L.

Consider the calculation of P_L based on Figure 6.10. Here P_L and P_H are defined with respect to the noninverting output to avoid possible ambiguity created by complementary outputs. Therefore, P_L corresponds to the condition in which Q_R is conducting and Q_I is cut off. In this situation,

$$\begin{aligned} P_L &= \left(-V_{EE}\right)\left(-I_{EEL}\right) = V_{EE}\left[I_{RE} + I_1 + I_2\right] \\ &= V_{EE}\left[\frac{V_{REF} - V_{BEA} - \left(-V_{EE}\right)}{R_E} + \frac{V_{OH} - \left(-V_{EE}\right)}{R_O} + \frac{V_{OL} - \left(-V_{EE}\right)}{R_O}\right] \\ &= V_{EE}\left[\frac{V_{REF} - V_{BEA} + V_{EE}}{R_E} + \frac{V_{OH} + V_{OL} + 2V_{EE}}{R_O}\right] \end{aligned} \tag{6.25}$$

Now consider the calculation of P_H based on Figure 6.11. P_H is defined with respect to the noninverting output also. Then, with Q_I conducting,

$$\begin{aligned} P_H &= \left(-V_{EE}\right)\left(-I_{EEH}\right) = V_{EE}\left[I_{RE} + I_1 + I_2\right] \\ &= V_{EE}\left[\frac{V_{OH} - V_{BEA} - \left(-V_{EE}\right)}{R_E} + \frac{V_{OL} - \left(-V_{EE}\right)}{R_O} + \frac{V_{OH} - \left(-V_{EE}\right)}{R_O}\right] \\ &= V_{EE}\left[\frac{V_{OH} - V_{BEA} + V_{EE}}{R_E} + \frac{V_{OH} + V_{OL} + 2V_{EE}}{R_O}\right] \end{aligned} \tag{6.26}$$

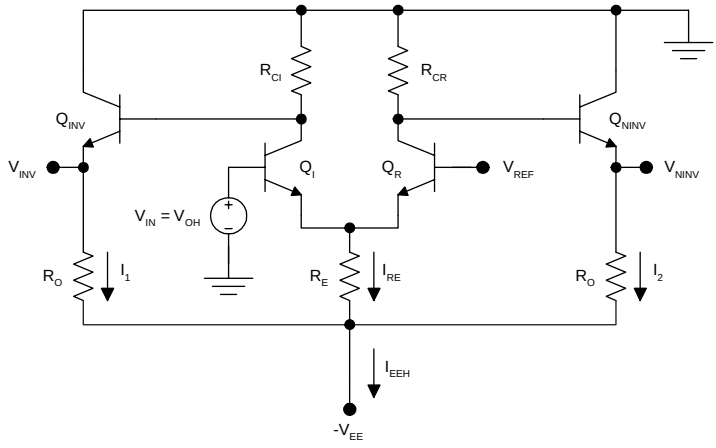


FIGURE 6.11
ECL gate for determination of P_H.

The average DC dissipation is the average of P_L and P_H:

$$P_{DC} = \frac{P_L + P_H}{2} = V_{EE}\left[\frac{V_{OH} + 2V_{REF} - V_{BEA} + 2V_{EE}}{R_E} + \frac{V_{OH} + V_{OL} + 2V_{EE}}{R_O}\right]. \quad (6.27)$$

It is important to notice that the power scales in inverse fashion with the resistors in the circuit. Also, because the dissipation involves square terms in V_{EE}, changing the supply voltage will have a strong impact on the dissipation. The connection of fan-out gates has no effect on the average power as defined here because the additional current flows to the V_{EE} connections for the fan-out gates rather than the loaded gate.

Example 6.3

Calculate P_L, P_H, and P_{DC} for the ECL gate depicted in Figure 6.12.

Solution. With Q_R conducting and V_{OL} approximated as –2 V_{BEA}, the power dissipation is

$$P_L = V_{EE}\left[\frac{V_{REF} - V_{BEA} + V_{EE}}{R_E} + \frac{V_{OH} + V_{OL} + 2V_{EE}}{R_O}\right]$$

$$= 5.2\text{ V}\left[\frac{-1.175\text{ V} - 0.75\text{ V} + 5.2\text{ V}}{0.78\text{ k}\Omega} + \frac{-0.75\text{ V} - 1.5\text{ V} + 10.4\text{ V}}{2\text{ k}\Omega}\right] = 44\text{ mW}$$

With Q_I conducting, and V_{OL} approximated as –2 V_{BEA},

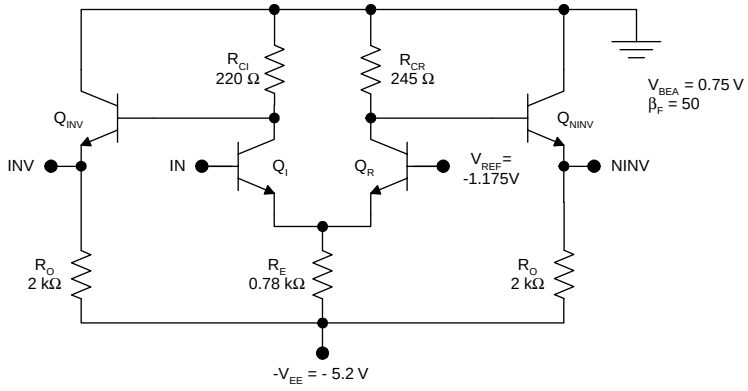


FIGURE 6.12
Example of ECL gate for calculation of P_{DC}.

$$P_H = V_{EE}\left[\frac{V_{OH} - V_{BEA} + V_{EE}}{R_E} + \frac{V_{OH} + V_{OL} + 2V_{EE}}{R_O}\right]$$

$$= 5.2\text{ V}\left[\frac{-0.75\text{ V} - 0.75\text{ V} + 5.2\text{ V}}{0.78\text{ k}\Omega} + \frac{-0.75\text{ V} - 1.5\text{ V} + 10.4\text{ V}}{2\text{ k}\Omega}\right] = 46\text{ mW}$$

The average DC dissipation is therefore

$$P_{DC} = \frac{P_L + P_H}{2} = \frac{44\text{ mW} + 46\text{ mW}}{2} = 45\text{ mW}.$$

6.6 Propagation Delays

ECL circuits are the fastest commercially available gates with respect to off-chip data rates. Propagation delays of less than 25 ps have been obtained even with off-chip loads. This compares favorably with the *on-chip* gate delays achievable by CMOS.

Underlying the speed of ECL are three design advantages: avoidance of saturated transistors, use of small logic swings, and use of emitter follower output drivers. The first two considerations make ECL gates fast internally whereas the latter makes ECL gates less susceptible to loading effects. Determination of the transient response in ECL circuits is rather complex and best done using computer simulations.[11–14] However, the propagation delays can be estimated using simple design calculations for two relevant situations: 1) the unloaded case and 2) the case of a lumped RC load. These limiting

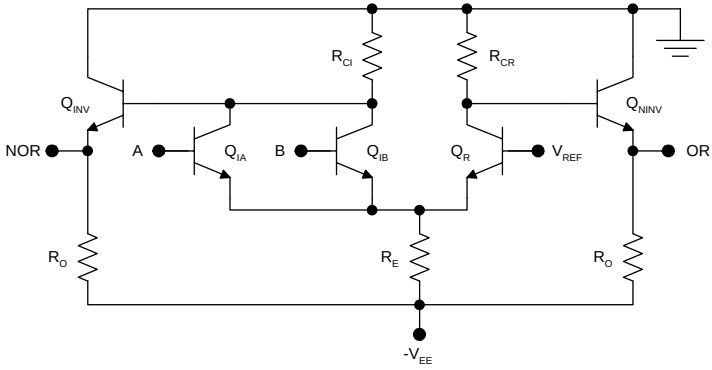


FIGURE 6.13
Unloaded ECL OR–NOR gate for consideration of the propagation delay.

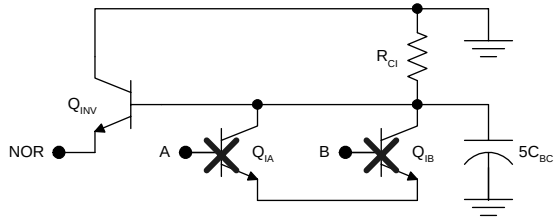


FIGURE 6.14
Simplified ECL circuit for consideration of the propagation delay.

cases and their associated design equations provide an excellent starting point from which physical ECL gate designs can be readily refined using computer tools.

6.6.1 Unloaded Case

Consider the unloaded ECL two-input OR–NOR gate as shown in Figure 6.13. Suppose V_{INA} is held at logic zero and a high-to-low transition is applied at V_{INB}. The switching speed is limited by how fast the voltage at the base of the emitter follower Q_{INV} can be brought to zero. Assuming that Q_{IB} turns off abruptly, the problem simplifies to a first-order RC circuit as shown in Figure 6.14.

The relevant resistance is R_{CI}. The total capacitance loading the node under consideration is

$$C = C_{BC}(Q_{INV}) + C_{BC}(Q_{IA}) + C_{BC}(Q_{IB}) + C_{CS}(Q_{IA}) + C_{CS}(Q_{IB}) \approx 5C_{BC}\,, \quad (6.28)$$

where C_{BC} denotes the base–collector capacitance and C_{CS} denotes the collector–substrate capacitance. Using the preceding approximation

$$V_B(Q_{INV}) \approx -V_{BEA} \exp\left(-\frac{t}{5R_{CI}C_{BC}}\right). \tag{6.29}$$

Using the 50% definition for the propagation delay,

$$t_{PLH} \approx 5R_{CI}C_{BC} \ln(2). \tag{6.30}$$

Furthermore, the switching performance is symmetric in the unloaded case, so

$$t_{PLH} \approx t_{PHL} \approx t_P \approx 5R_C C_{BC} \ln(2). \tag{6.31}$$

This is a very rough approximation because it ignores second-order effects and the voltage dependence of the parasitic capacitances. However, it is also very useful because it shows that the propagation delay is directly proportional to R_C, which indicates a direct trade-off between speed and power if all resistors in the circuit scale together. In addition, the propagation delays are directly proportional to the parasitic capacitances. Therefore, if the power is not to be scaled, improved dynamic performance requires the physical redesign of the bipolar transistors to achieve smaller parasitic capacitances.

6.6.2 Lumped RC Load

Consider the ECL two-input OR–NOR gate with a lumped RC load as shown in Figure 6.15. The load is considered to comprise a resistance R_L in parallel with a capacitance C_L applied to the inverting (NOR) output. With the RC load connected, the low-to-high propagation delay is increased slightly from the unloaded case to

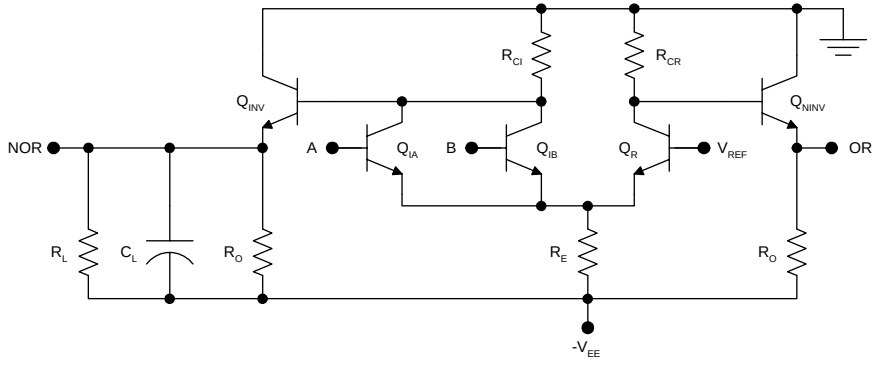


FIGURE 6.15
ECL OR–NOR gate with lumped RC load for consideration of the propagation delay.

$$t_{PLH} \approx 5R_C C_{BC} \ln(2) + \left[\left(\frac{R_{CI}}{\beta_F} + g_m\right) \| R_O \| R_C\right] C_L \ln(2)$$
$$\approx 5R_C C_{BC} \ln(2) + g_m C_L \ln(2) \tag{6.32}$$

where g_m is the transconductance of the emitter follower. The emitter follower at the output conducts throughout the low-to-high transition, providing excellent current drive to charge the load capacitance. The output impedance of the emitter follower ($\sim g_m$) is typically less than 10 Ω. In the case of the high-to-low transition, however, the emitter follower at the output may turn off, or at least decrease, its emitter current. Then the propagation delay becomes limited by the external RC product:

$$t_{PHL} \approx \left(R_O \| R_C\right) C_L \ln(2) . \tag{6.33}$$

This reveals an important deficiency in standard ECL circuit designs: the lack of active pull-down. To address this issue, ECL circuits with active pull-down (APD–ECL) have been developed to allow short propagation delays with a wide range of lumped capacitive or RC loads.

6.7 Logic Design

With complementary outputs, the basic ECL gate shown in Figure 6.16 is an inverter–buffer. An OR–NOR gate is constructed by paralleling two or more input transistors. The gate shown in Figure 6.17 is a two-input OR–NOR gate. In this circuit, if either input goes high, the associated input transistor turns on and brings the inverting output low. This causes Q_R to turn off, bringing the noninverting output high. If both inputs are low, then both input transistors are cut off but Q_R conducts. This is the only condition under which the inverting output goes high and the noninverting output goes low; therefore, the inverting output provides the NOR function while the noninverting output provides the OR function.

The AND and NAND functions are implemented in ECL using the series gating approach (the input transistors are placed in series). The two-input AND–NAND gate shown in Figure 6.18 illustrates this approach. If both inputs go high, the series input transistors Q_{IA} and Q_{IB} turn on and bring the NAND output low. If either input goes low, the associated input transistor is cut off but the associated reference transistor turns on, bringing the AND output low.

Two reference voltages are required for a two-input series-gated AND–NAND ECL gate. The most positive reference voltage is the same as that used in ECL OR–NOR gates:

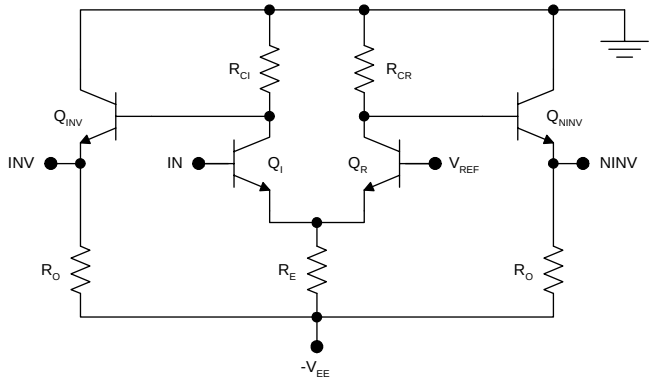


FIGURE 6.16
ECL inverter buffer.

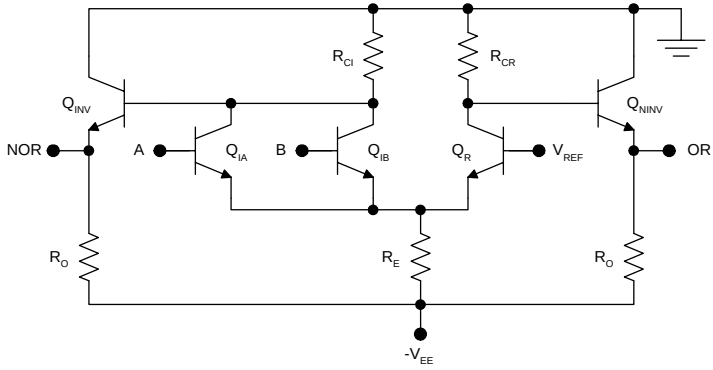


FIGURE 6.17
ECL OR2–NOR2 gate.

$$V_{REF1} \approx -\frac{3V_{BEA}}{2}. \tag{6.34}$$

The second reference voltage must be lowered by one base–emitter drop in order to prevent saturation of Q_{IB} or Q_{R2}. Thus,

$$V_{REF2} \approx -\frac{5V_{BEA}}{2}. \tag{6.35}$$

Also, input B must be level shifted by one base–emitter drop, which is achieved using the emitter follower Q_{LB}.

Other more complicated logic functions may be implemented using series gating in conjunction with paralleled input transistors. An example with six inputs is shown in Figure 6.19. The functions implemented by this circuit are:

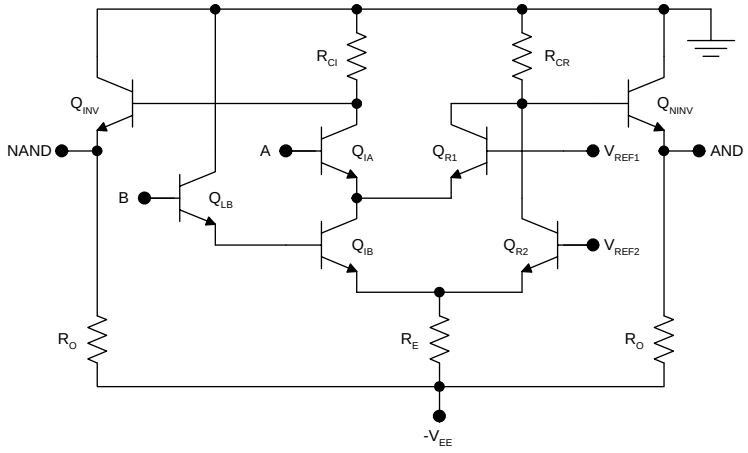


FIGURE 6.18
ECL AND2-NAND2 circuit using the series-gating approach.

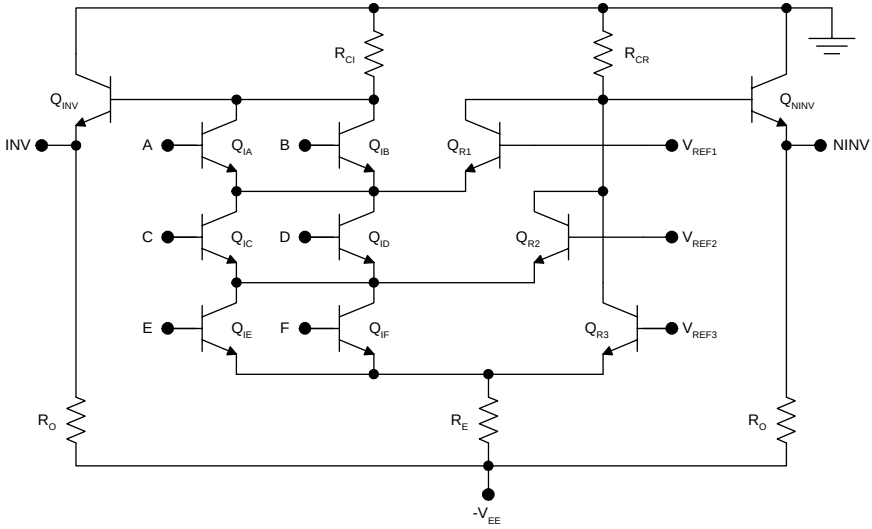


FIGURE 6.19
Implementation of complex logic functions in ECL using series gating in conjunction with paralleled input transistors.

$$Y = (A + B)(C + D)(E + F) \tag{6.36}$$

and

$$\overline{Y} = \overline{(A + B)(C + D)(E + F)}\,. \tag{6.37}$$

As a consequence of the series gating, inputs *C* and *D* must be level shifted in the negative direction by one base–emitter drop. (This is not shown in the figure for the sake of simplicity.) Also, inputs E and F must be level shifted by two p–n junction drops in the negative direction. V_{REF3} must be one base–emitter drop more negative than V_{REF2}, which is one base–emitter drop more negative than V_{REF1}.

More complicated logic functions may be implemented. However, if saturation is to be prevented in all reference transistors, the maximum number of stages that can be series gated is given by

$$S_{MAX} \leq \frac{|V_{EE}| - (V_{OH} - V_{OL}) - V_{CS}}{V_{BEA}}, \tag{6.38}$$

where V_{CS} is the voltage drop across the current source and it is assumed that the voltage across each stage is designed to be V_{BEA} to prevent saturation of any of the transistors. Based on this design approach, the standard supply voltage of –5.2 V allows the series gating of a maximum of three stages. This makes the implementation of complex logic functions in ECL relatively cumbersome and complicates the implementation of low-voltage ECL gates as well.

6.8 Temperature Effects in ECL

Temperature effects are more important in ECL than in any other logic family used today because the output levels depend on the value of V_{BEA}, which varies with temperature. This temperature variation is exaggerated by the small logic swing inherent in ECL circuitry.

Consider the variation of the base–emitter voltage in a bipolar transistor under constant current conditions. The temperature variation is given by

$$\left.\frac{dV_{BEA}}{dT}\right|_{\text{I=constant}} = \frac{d\left[(kT/q)\ln(I/I_S)\right]}{dT} = \frac{V_{BEA}}{T} - \frac{kT}{q}\left(\frac{1}{I_S}\frac{dI_S}{dT}\right), \tag{6.39}$$

where I_S is the saturation current for the transistor. Typically, the value of this temperature derivative is –2 mV/°C. The negative sign indicates that V_{BEA} decreases with increasing temperature.

The ECL circuit of Figure 6.16 is not temperature compensated. The temperature variation of the output high voltage in such a circuit is

$$\frac{dV_{OH}}{dT} = -\frac{dV_{BEA}}{dT} \approx 2 \text{ mV/°C} \tag{6.40}$$

and the temperature variation of the output low voltage is given by

$$\frac{dV_{OL}}{dT} = -\frac{R_{CI}}{R_E}\left(\frac{dV_{OH}}{dT} - \frac{dV_{BEA}}{dT}\right) - \frac{dV_{BEA}}{dT} \approx 1 \text{ mV/°C}. \tag{6.41}$$

Thus output levels move closer to ground with an increase in temperature. With a fixed reference voltage, uncompensated ECL will not function properly above 60°C.

In order to remain centered between V_{OL} and V_{OH}, the reference voltage should have a temperature variation of +1.5 mV/°C. The ECL II, ECL 10k, and ECL III families were designed with a bias driver that has this temperature dependence (described in the next section). Although this is superior to the uncompensated design, it is still not entirely satisfactory because a large system may contain circuits at different temperatures. Therefore, the ideal solution is to make the output levels independent of temperature — an approach taken in designing ECL 100k and all other modern ECL circuits.

6.9 ECL Circuit Families

The first family of emitter-coupled logic was ECL I, introduced by Motorola.[15] This family is based on the circuit diagram of Figure 6.20. Two V_{CC} lines are grounded on the circuit board. V_{CC1} is the "dirty ground"; fast switching transients on the outputs are accompanied by current spikes in the emitter followers, resulting in ohmic voltage spikes on the dirty ground.[16] V_{CC2} is called the "clean ground" because the nearly constant current flowing in this line makes it electrically quiet. ECL I was superseded by ECL II, ECL III, and ECL 10k. All three families are similar and based on the circuit diagram shown in Figure 6.21.[17,18]

ECL II, ECL 10k, and ECL III have a number of important similarities, e.g., they use a temperature-compensated bias driver (comprising Q_B, D_{L1}, D_{L2}, R_{BE}, R_{BL}, and R_{BH}). The key differences among the three families are in the resistor values and the transistor designs. Table 6.2 summarizes the circuit values and performance for these families.

ECL 100k[17,18] exhibits temperature coefficients superior to those of the other families due to an improved bias driver and a modified current switch. In addition, the improved bias driver makes ECL 100k less sensitive to supply voltage variations. Finally, ECL 100k exhibits improved speed and power performance because of the use of a reduced supply voltage and transistors with reduced parasitics.

An ECL 100k inverter buffer is shown in Figure 6.22. This circuit was designed with two significant advantages over the earlier ECL families. First, the emitter resistor has been replaced by a BJT current source, which makes the current in the switch nearly independent of the supply voltage. Therefore,

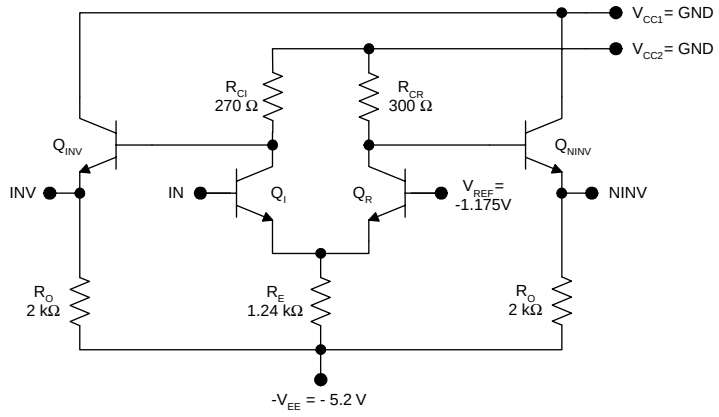


FIGURE 6.20
ECL I inverter–buffer.

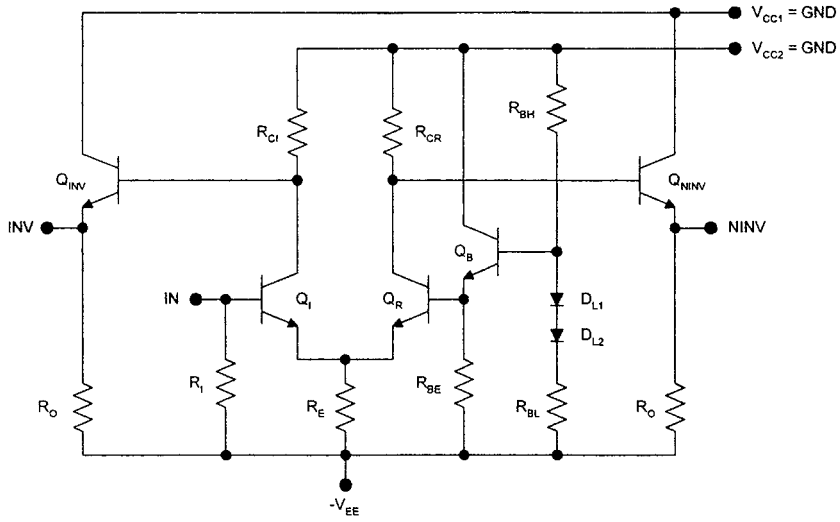


FIGURE 6.21
Circuit diagram for the ECL II, ECL 10k, and ECL III families.

the voltage transfer characteristic is practically independent of the supply voltage — a requirement in large systems with multiple power supplies. Second, modifying the current switch by adding the two diodes and R_C allows superior temperature compensation. (The temperature derivatives are less than 0.1 V/°C for V_{OL} and V_{OH}.)

A constant current flows in the current source transistor Q_E:

$$I_{EE} = \frac{V_{REF2} - V_{BEA} - (-V_{EE})}{R_E}. \tag{6.42}$$

TABLE 6.2

Circuit Values and Performance for ECL II, ECL 10k, and ECL III[a]

Symbol	ECL II	ECL 10k	ECL III
R_E	1.18	0.78	0.365
R_{CI}	0.29	0.22	0.100
R_{CR}	0.30	0.245	0.112
R_I	—	50	50
R_O	2.0	—	—
R_{BH}	0.30	0.91	0.35
R_{BL}	2.3	5.0	1.96
R_{BE}	2.0	6.1	2.0
$-V_{EE}$ (V)	–5.2	–5.2	–5.2
P (mW)	20	25	60
t_p(ns)	4	2	1

[a] Resistor values are in kΩ.

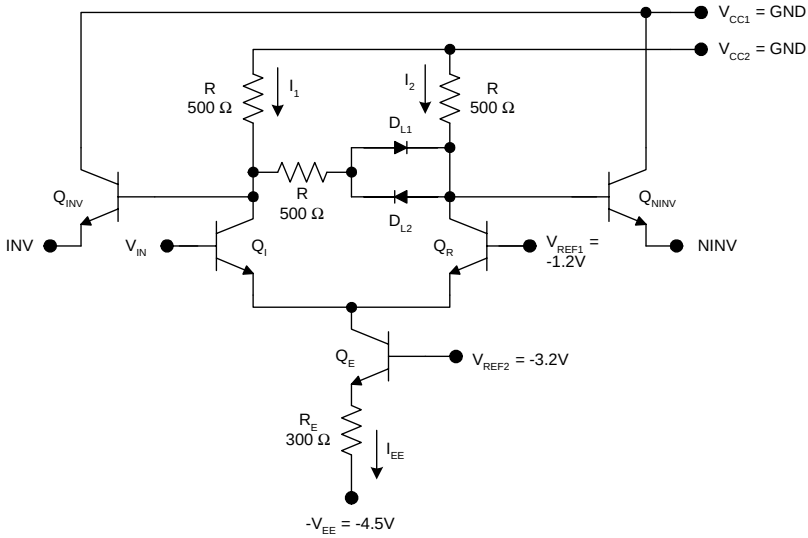


FIGURE 6.22
ECL 100k inverter–buffer gate.

With Q_R conducting, D_{L1} conducts and V_{NINV} goes low. Under these conditions,

$$I_1 = \frac{-\left(V_{OL} + V_D + V_{BEA}\right)}{2R} \tag{6.43}$$

and

$$I_2 = \frac{-\left(V_{OL} + V_{BEA}\right)}{R}. \tag{6.44}$$

Because

$$I_1 + I_2 = I_{EE}, \tag{6.45}$$

$$V_{OL} = -\frac{2RI_{EE}}{3} - \frac{V_D}{3} - V_{BEA}. \tag{6.46}$$

With Q_1 conducting, D_{L2} conducts and V_{NINV} goes high. Under these conditions,

$$I_1 = \frac{-(2V_{OH} + 2V_{BEA}) + V_D}{R} \tag{6.47}$$

and

$$I_2 = \frac{-(V_{OH} + V_{BEA})}{R}. \tag{6.48}$$

Because

$$I_1 + I_2 = I_{EE}, \tag{6.49}$$

$$V_{OH} = -\frac{RI_{EE}}{3} + \frac{V_D}{3} - V_{BEA}. \tag{6.50}$$

The ECL 100k bias driver is shown in Figure 6.23. Unlike the earlier families, ECL 100k uses two reference voltages. The first is applied to the reference transistor of the current switch as was done before. The second reference voltage is applied to the base of the current source transistor. The essential function of the bias driver is as follows.

Q_{S2} is a temperature-compensated current source,

$$I_{CB3} \approx \frac{V_{DL} - V_{BES2}}{R_{E2}}. \tag{6.51}$$

This maintains a temperature-compensated V_{REF1}:

$$V_{REF1} \approx -I_{CB3}R_{BH} - V_{BEA}. \tag{6.52}$$

The shunt regulator, Q_{SH}, maintains a constant voltage across R_2, which regulates the current source Q_{S1}:

$$I_{CS1} = \frac{V_{BESH}}{R_2}, \tag{6.53}$$

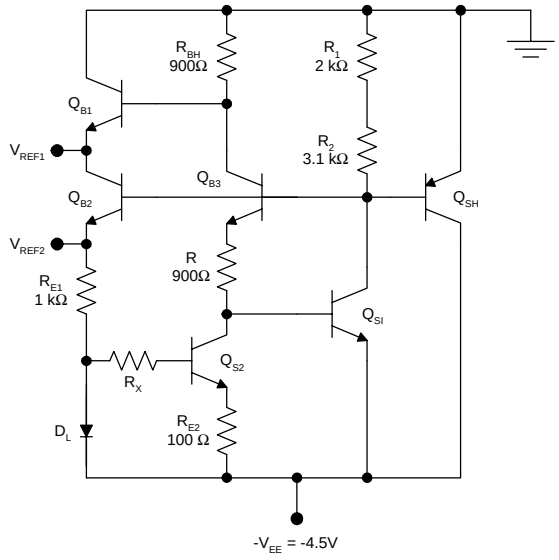


FIGURE 6.23
ECL 100k bias driver.

which serves to regulate V_{REF2}:

$$V_{REF2} \approx -V_{EE} + V_{BEA} + I_{CB3}R \,. \tag{6.54}$$

It is important to appreciate that V_{REF1} and V_{REF2} are temperature compensated; furthermore, V_{REF2} shifts up or down with V_{EE}. This dependence keeps the current I_{EE} in the current switch independent of the supply voltage as desired.

Other families of ECL have emerged since ECL 100k; most use similar circuit designs. In most cases the important differences are in the transistor fabrication (oxide isolation, polysilicon emitters) and scaling of the resistors. These modifications have resulted in circuits with propagation delays under 25 ps and power delay products less than 100 fJ. Further improvements in performance have been achieved using ECL circuits with active pull-down (discussed in the next section). ECL circuits have also been adapted to positive voltage applications.[27,28]

6.10 Active Pull-Down ECL (APD ECL)

An important speed limitation with conventional ECL circuits is observed in the high-to-low propagation delay. With a large capacitive load, the emitter follower at the output can turn off so that t_{PHL} becomes limited by the external

RC time constant. This problem may be addressed using an active pull-down circuit in which an extra bipolar transistor is included to discharge the load capacitance rapidly during high-to-low transitions.[19–26]

Design of active pull-down ECL (APD ECL) has two approaches. The first uses a coupling capacitor to drive the pull-down transistor and is called AC-coupled active pull-down ECL (AC-APD ECL); the second is level-sensitive active pull-down ECL (LS-APD ECL). The LS-APD ECL approach is more versatile because the pull-down drive self-adjusts to a wide range of load capacitance.

6.10.1 AC-Coupled Active Pull-Down ECL (AC-APD ECL)

In the AC-APD ECL circuit, a capacitor is used to provide base drive for the pull-down transistor. This is illustrated in the inverter of Figure 6.24, in which a displacement current flows in C_X during the output high-to-low transition and turns on the pull-down transistor, Q_{PD}.

A challenge in designing such a circuit is that the coupling capacitor, C_X, must be appropriately sized for the load capacitance. If C_X is too large for the particular load capacitance, excessive simultaneous conduction of the pull-up and pull-down transistors occurs, resulting in wasted power. Also, the output voltage will overshoot. On the other hand, if C_X is too small for a particular load capacitance, then the pull-down transistor will turn off prematurely and t_{PHL} will be degraded. This need to tailor the coupling capacitor makes the AC-coupled approach somewhat difficult to implement. As a consequence, level-sensitive active pull-down is the preferred circuit approach.

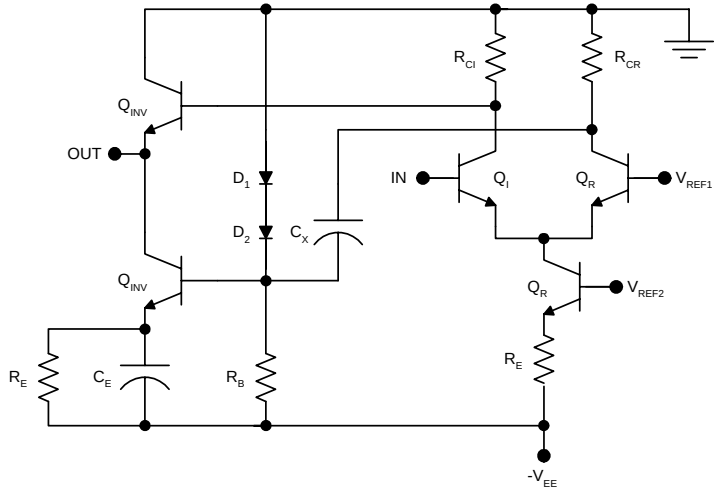


FIGURE 6.24
AC-APD ECL inverter circuit.

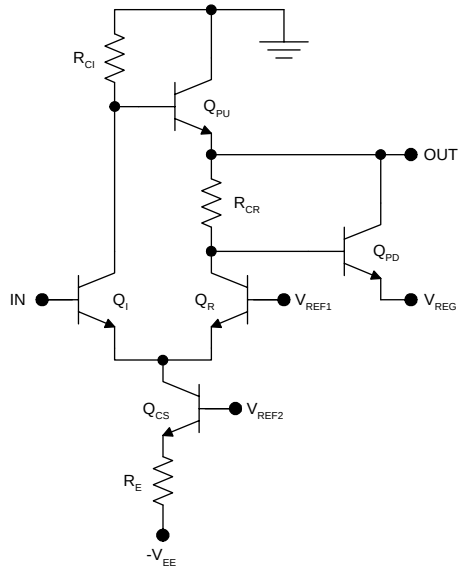


FIGURE 6.25
LS-APD ECL inverter circuit.

6.10.2 Level-Sensitive Active Pull-Down ECL (LS-APD ECL)

The advantage of the level-sensitive active pull-down ECL (LS-APD ECL) circuit is that it is self-adjusting. That is, the pull-down drive automatically adapts to the capacitive loading at the output, which ensures optimum dynamic performance without excess power dissipation. An example of an LS-APD ECL inverter is shown in Figure 6.25, in which V_{REG} is set to

$$V_{REG} = V_{OL} - V_{BEA}\,. \tag{6.55}$$

Therefore, during an output high-to-low transition, the pull-up transistor Q_{PU} turns off and the discharging of the load provides base drive to the pull-down transistor, Q_{PD}. The pull-down transistor will not turn off until the output voltage reaches V_{OL}. Therefore, the level-sensitive circuit is self-adjusting with respect to the load. The only disadvantage of the LS-APD ECL circuit is that it precludes the use of complementary outputs.

6.11 Low-Voltage ECL (LV-ECL)

In recent times much emphasis has been placed on the reduction of supply voltages in digital circuits. In the case of ECL, reduction of the supply voltage

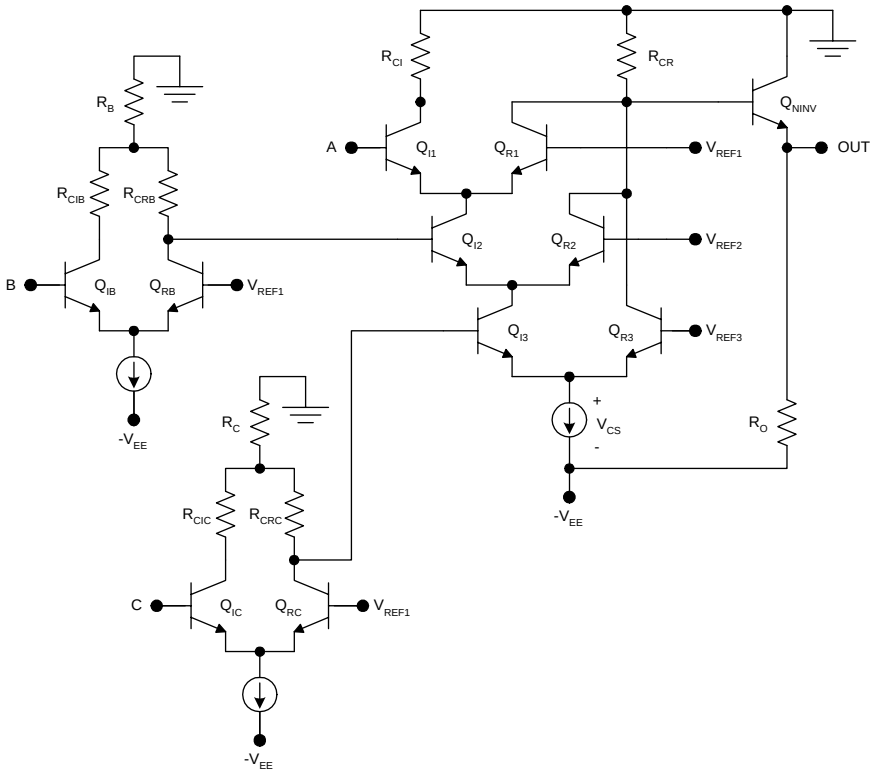


FIGURE 6.26
Low-voltage ECL (LV-ECL) AND3 circuit utilizing three stages of series-gated transistors.

not only reduces dissipation but also eases problems of interfacing with low-voltage CMOS.

The most significant challenge in the design of low-voltage ECL has to do with series-gated circuits. In conventional ECL circuits, each successive stage in a series-gated stack drops one base–emitter drop. This prevents the saturation of any of the transistors because the collector–emitter voltage in any conducting transistor in the stack will be equal to V_{BEA}. Also, this scheme allows for convenient level shifting of the input signals because the nth stage of the series-gated stack requires the input signals to be level shifted by nV_{BEA}.

In low-voltage ECL (LV-ECL),[29–32] voltage drops across the individual stages in the series-gated stack are reduced. It is still possible to prevent saturation if each stage drops $V_{BEA}/2$. On the other hand, the use of a fractional base–emitter drop across each stage does not allow the simple level-shifting scheme used in conventional ECL. Instead, the required level shifting is achieved using current mode logic subcircuits as shown in Figure 6.26.

In the LV-ECL AND3 gate circuit of Figure 6.26, input A is applied directly to the first stage. Input B is level shifted by the current mode logic circuit

comprising Q_{IB} and Q_{RB}. In this current mode level shifter, the resistor R_B is chosen so that input B is level shifted down by $V_{BEA}/2$. In similar fashion, input C is level shifted down by V_{BEA} using a second current mode logic circuit.

Each conducting stage of the series-gated AND gate drops $V_{BEA}/2$. If the logic swing is scaled down to $V_{BEA}/2$ and the drop across the current source is V_{BEA}, then the circuit can operate with $|V_{EE}| = 3\ V_{BEA}$. Practical LV-ECL circuits allow three stages of series gating with a –2 V supply.[32] Even lower voltage circuits can be achieved if the logic swings are further reduced or if SiGe transistors with lower turn-on voltages are used.

6.12 PSPICE Simulations

Simulations were performed using PSPICE 9.1, which can be downloaded free.[33] The BJT model parameters used in all simulations are provided in Table 6.3.

6.12.1 Voltage Transfer Characteristics

The voltage transfer characteristics were simulated for the *unloaded* ECL gate of Figure 6.27; the results appear in Figure 6.28. It can be seen that the transition widths are close to the 0.1 V rule-of-thumb value. The "hook" in the inverting characteristic is the result of saturation in the input transistor, the onset of which occurs at $V_{IN} \approx -0.2$ V.

TABLE 6.3

BJT SPICE Parameters

Parameter	Value	Units
IS	2.0f	A
BF	70	—
NF	1	—
BR	0.5	—
NR	1	—
CJE	0.3p	F
VJE	0.8	V
MJE	0.333	—
TF	0.05n	s
CJC	0.15p	F
VJC	0.75	V
MJC	0.5	—
TR	0.5n	s

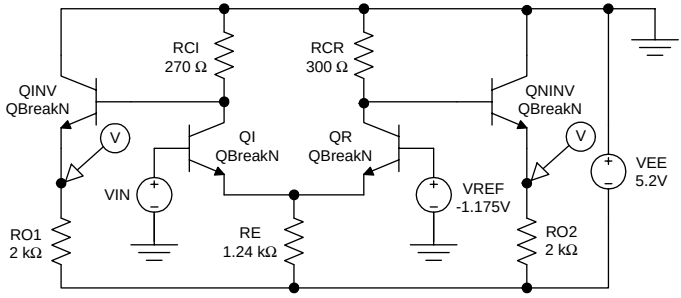


FIGURE 6.27
ECL circuit used for the simulation of the VTCs.

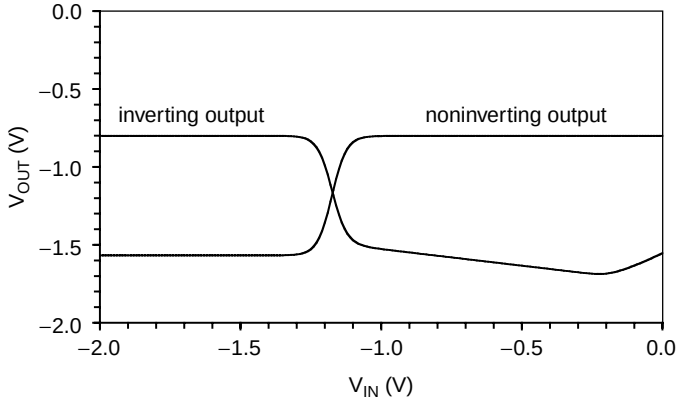


FIGURE 6.28
Simulated VTCs for the ECL circuit.

6.12.2 Temperature Effects

The effect of the temperature was simulated using a nested temperature sweep within the DC sweep. The resulting transfer characteristics appear in Figure 6.29. As expected, V_{OL} and V_{OH} become more positive as the temperature increases. It can also be seen that the point of saturation in the input transistor is a function of temperature.

6.12.3 Propagation Delays

The propagation delays for an *unloaded* ECL inverter were determined using the circuit of Figure 6.30.

The results of the transient simulation for the noninverting output appear in Figure 6.31. With no load, the propagation delays are symmetric: t_{PHL} = t_{PLH} = 48 ns. Although ECL circuitry is inherently fast, t_{PHL} deteriorates if a

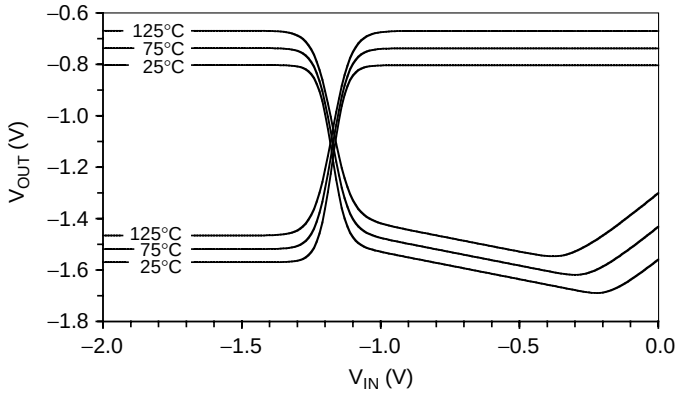


FIGURE 6.29
Simulated VTCs for the ECL circuit, with the temperature as a parameter.

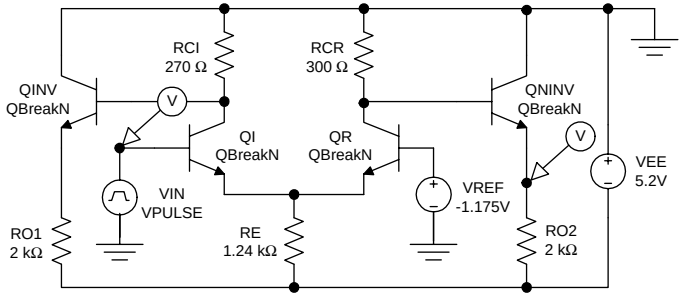


FIGURE 6.30
ECL circuit for simulation of unloaded propagation delays for the noninverting output.

large load capacitance is applied at the output. This was explored using transient simulations for the circuit of Figure 6.32.

The results of transient simulations in Figure 6.33 show that the high-to-low transition slows considerably if the emitter follower turns off because then t_{PHL} becomes limited by the $R_{O2}C_L$ time constant. This behavior is best illustrated in Figure 6.34, which shows the propagation delays as a function of the load capacitance. Both propagation delays increase monotonically with the load capacitance; however, the slopes are very different:

$$\frac{\partial t_{PHL}}{\partial C_L} \approx \frac{2.73\text{ ns} - 0.162\text{ ns}}{15\text{ pF} - 1\text{ pF}} = 183\Omega\,,$$

but

$$\frac{\partial t_{PLH}}{\partial C_L} \approx \frac{0.53\text{ ns} - 0.130\text{ ns}}{15\text{ pF} - 1\text{ pF}} = 29\Omega\,.$$

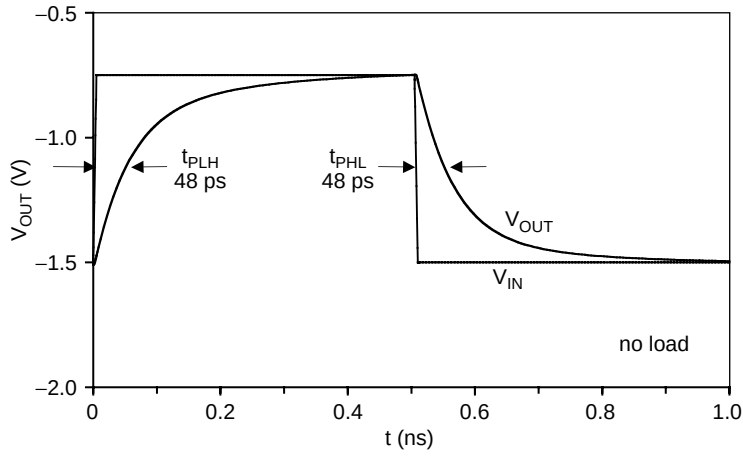


FIGURE 6.31
Simulated ECL transient response with no load.

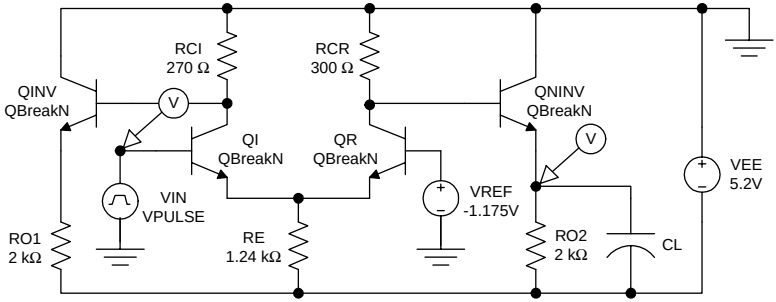


FIGURE 6.32
ECL circuit for the simulation of the propagation delays with a capacitive load.

6.12.4 Level-Sensitive Active Pull-Down ECL

The characteristics of LS-APD ECL were simulated for comparison to standard ECL circuits. The voltage transfer characteristic for the *unloaded* LS-APD ECL circuit of Figure 6.35 was simulated using a DC sweep. The results of Figure 6.36 show that V_{OL} = –1.50 V and V_{OH} = –0.77 V. Complementary outputs are not available for the LS-APD circuit so only an inverting characteristic is obtained. The important advantage of LS-APD ECL compared to conventional ECL circuitry is that t_{PHL} is less sensitive to the load capacitance. This can be shown using transient simulations for the *loaded* circuit of Figure 6.37.

The transient simulation of Figure 6.38 for the case of C_L = 15 pF shows that the transient response is approximately symmetric: $t_{PLH} \approx t_{PHL} \approx 500$ ps.

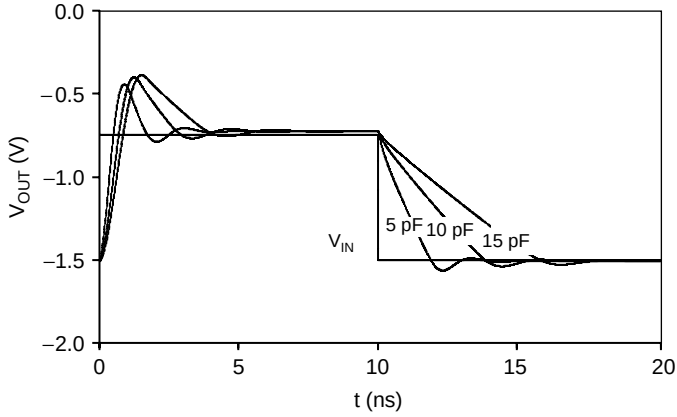


FIGURE 6.33
Simulated ECL transient response with the load capacitance as a parameter.

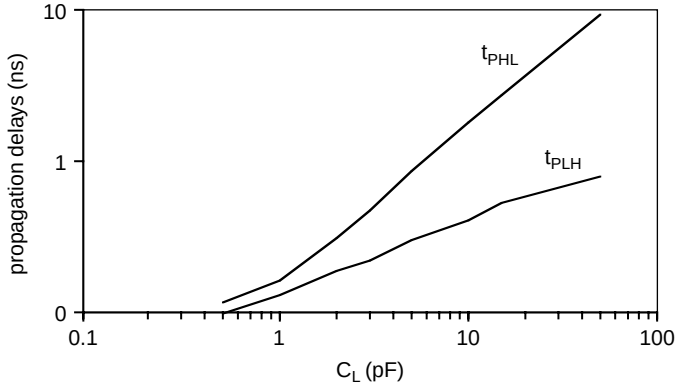


FIGURE 6.34
ECL propagation delays as a function of the load capacitance.

Additional transient simulations were performed to obtain the propagation delays as a function of the load capacitance. This behavior is best illustrated in Figure 6.39, which shows the propagation delays as a function of the load capacitance. Both propagation delays increase monotonically with the load capacitance, with similar slopes:

$$\frac{\partial t_{PHL}}{\partial C_L} \approx \frac{\partial t_{PLH}}{\partial C_L} \approx 30\Omega.$$

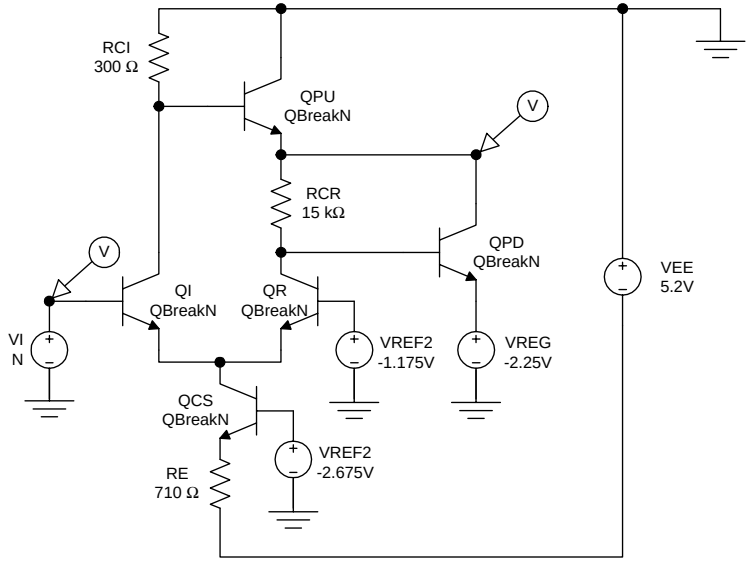


FIGURE 6.35
LS-APD ECL circuit used for simulation of the VTC.

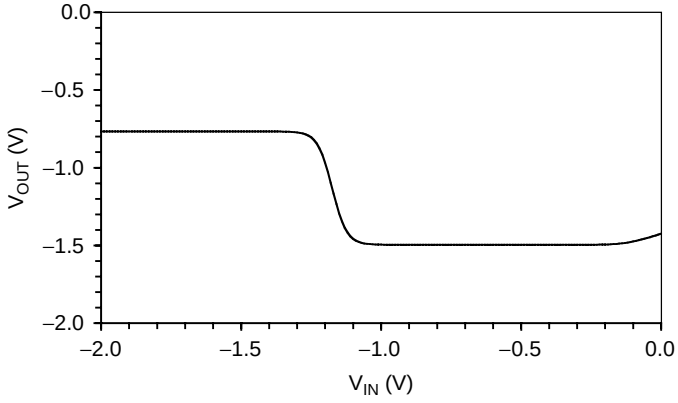


FIGURE 6.36
Simulated VTC for the LS-APD ECL circuit.

6.13 Summary

Emitter-coupled logic gates achieve the highest off-chip data rates of all silicon logic circuits. This speed is achieved in part by the use of current mode logic and avoidance of saturated bipolar transistors. In addition, ECL

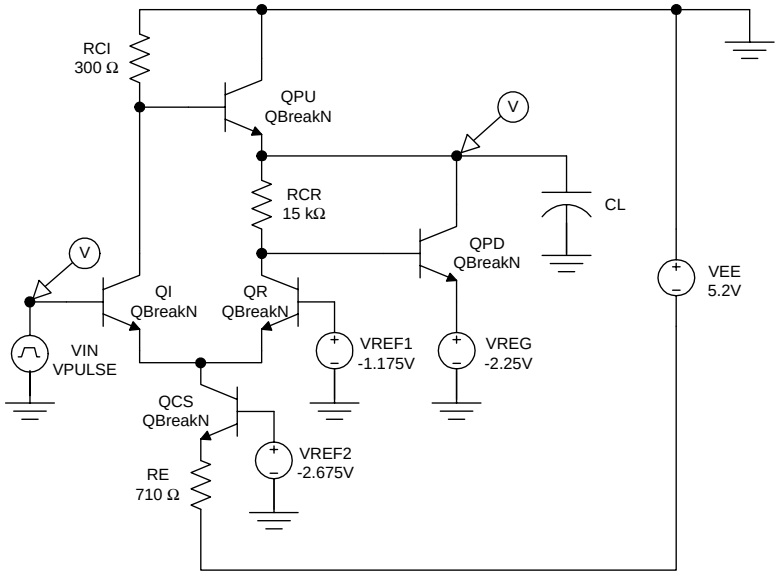


FIGURE 6.37
LS-APD ECL circuit for determination of the propagation delay with a capacitive load.

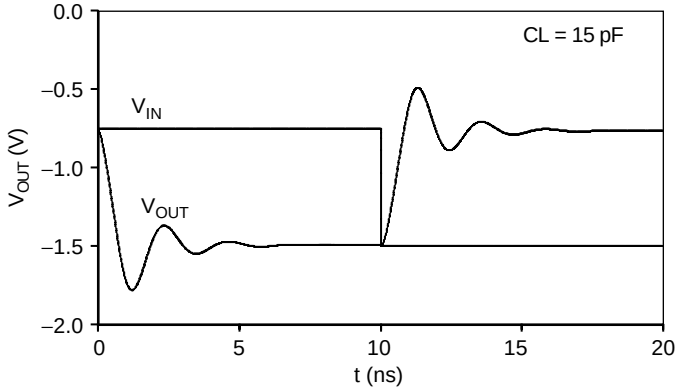


FIGURE 6.38
Simulated LS-APD ECL transient response with a 15-pF load capacitance.

gates use emitter follower outputs that make these gates less susceptible to capacitive loading at the output.

Temperature effects are important in ECL because the forward voltage for the base–emitter junction of a bipolar transistor decreases with increasing temperature. Sophisticated temperature compensation schemes are used to

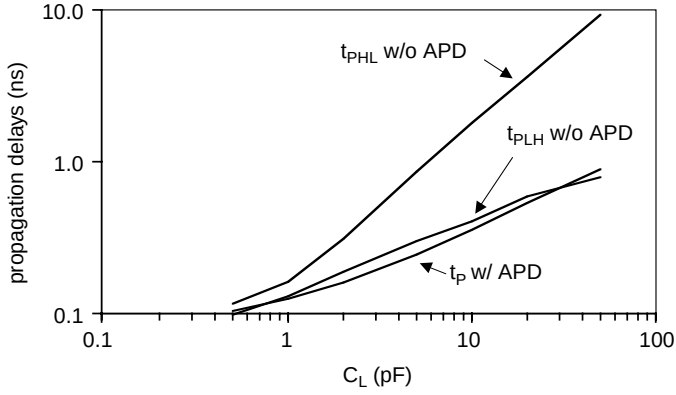


FIGURE 6.39
Propagation delays as a function of the load capacitance, for conventional ECL (without active pull-down) and level-sensitive active pull-down ECL.

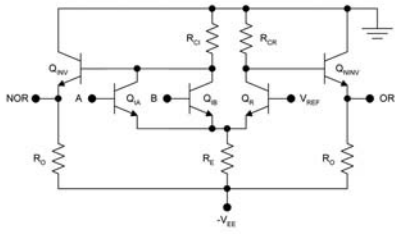
address this problem. ECL circuits with active pull-down are superior to passive pull-down ECL circuits when driving highly capacitive loads. These circuits, including AC-coupled active pull-down ECL (AC-AP ECL) and level-sensitive active pull-down ECL (LS-APD ECL) include a pull-down transistor at the output, greatly improving t_{PHL} for the case of a large load capacitance.

Recently, low-voltage ECL has been developed, enabling a drastic reduction in the power dissipation and also the power delay product. The primary challenge in the design of these circuits is the realization of the AND function. However, the reduction of the logic swing and the use of current mode logic level shifters have made it possible to realize three-input AND/NAND gates with a –2.0-V supply.

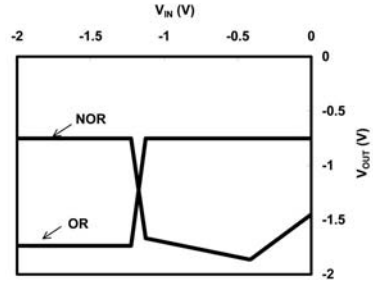
ECL circuitry is bound to remain important for high bitrate applications because of the higher transconductance values achieved with bipolar transistors compared to FETs. It is also quite likely that the implementation of SiGe bipolar transistors will further boost the performance of ECL circuitry over the next few years.

EMITTER-COUPLED LOGIC QUICK REFERENCE

ECL OR / NOR circuit



DC Voltage Transfer Characteristics



ECL, based on BJT current switches, is the fastest commercially available logic family. ECL gates achieve propagation delays in picoseconds with off-chip loads.

DC Voltage Transfer Characteristic

$$V_{IL} \approx V_{REF} - 0.05V \qquad V_{IH} \approx V_{REF} + 0.05V \qquad V_{OH} = -V_{BEA}$$

$$V_{OL} = \begin{cases} -R_{CI}\left(\dfrac{\beta_F}{1+\beta_F}\right)\dfrac{-2V_{BEA}-(-V_{EE})}{R_E} - V_{BEA}; & \text{inverting output} \\ -R_{CR}\left(\dfrac{\beta_F}{1+\beta_F}\right)\dfrac{V_{REF}-V_{BEA}-(-V_{EE})}{R_E} - V_{BEA}; & \text{non - inverting output} \end{cases}$$

DC Dissipation

$$P_H = V_{EE}\left[\frac{V_{OH}-V_{BEA}+V_{EE}}{R_E} + \frac{V_{OH}+V_{OL}+2V_{EE}}{R_O}\right]$$

$$P_L = V_{EE}\left[\frac{V_{REF}-V_{BEA}+V_{EE}}{R_E} + \frac{V_{OH}+V_{OL}+2V_{EE}}{R_O}\right]$$

Propagation Delays

$$t_{PLH} \approx t_{PHL} \approx t_P \approx 5R_C C_{BC} \ln(2)$$

Temperature Effects

$$\frac{dV_{OH}}{dT} = -\frac{dV_{BEA}}{dT} \qquad \frac{dV_{OL}}{dT} = -\frac{R_{CI}}{R_E}\left(\frac{dV_{OH}}{dT} - \frac{dV_{BEA}}{dT}\right) - \frac{dV_{BEA}}{dT}$$

Design Rules

$$V_{BEA} = 0.75V \text{ (ECL)} \qquad \frac{dV_{BEA}}{dT} \approx \frac{dV_D}{dT} \approx -2mV/^\circ C$$

f = 10^{-15} p = 10^{-12} n = 10^{-9} μ = 10^{-6} m = 10^{-3} k = 10^{3} M = 10^{6} G = 10^{9}

Laboratory Exercises

L6.1. Consider the *unloaded* ECL buffer–inverter circuit shown in Figure 6.40.

1. Determine and plot both voltage transfer characteristics using hand calculations.
2. Determine and plot both voltage transfer characteristics using SPICE.
3. Build the inverter and measure both voltage transfer characteristics using the x–y feature of an oscilloscope or virtual instrument and a low frequency input signal (1 kHz).
4. Plot the experimental, hand-calculated, and SPICE results together for comparison.

L6.2. Consider the *unloaded* ECL circuit of Figure 6.41 with *scaled resistors*. $1/4 < x < 4$.

1. Estimate P_H, P_L, and P_{DC} by hand calculations. Plot P_{DC} vs. x.
2. Determine P_H, P_L, and P_{DC} using SPICE. Plot P_{DC} vs. x.
3. Estimate the unloaded propagation delays using SPICE and plot the average propagation delay as a function of x.
4. Using the SPICE results, plot the power delay product vs. x.
5. Build the circuit and measure P_H, P_L, and P_{DC} as a function of x.

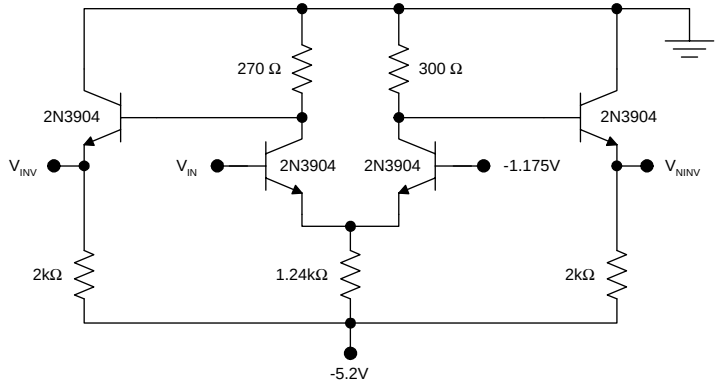


FIGURE 6.40
Unloaded ECL buffer–inverter circuit (L6.1).

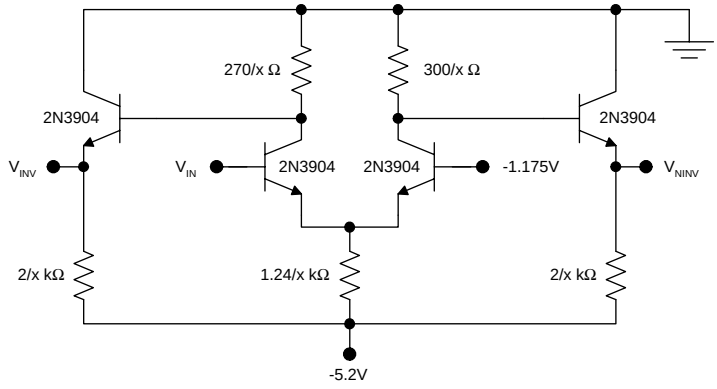


FIGURE 6.41
Unloaded ECL buffer–inverter circuit (L6.2).

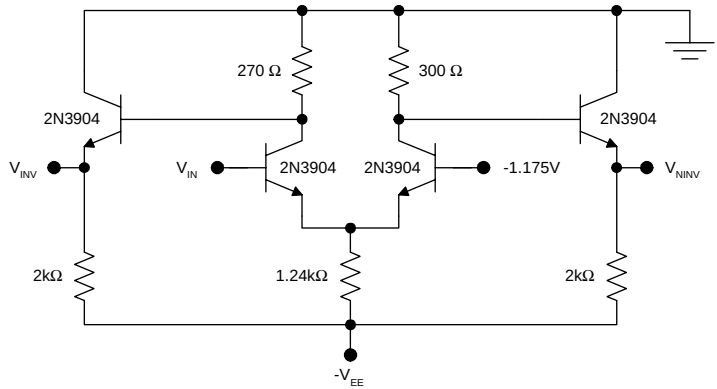


FIGURE 6.42
ECL circuit (L6.3).

L6.3. Consider the ECL circuit depicted in Figure 6.42. 4 V < V_{EE} < 6 V.

1. Determine and plot the voltage transfer characteristic using hand calculations. Plot the VTCs with V_{EE} as a parameter.
2. Determine and plot the voltage transfer characteristic using SPICE. Plot the VTCs with V_{EE} as a parameter.
3. Estimate P_H, P_L, and P_{DC} by hand calculations. Plot P_{DC} vs. V_{EE}.
4. Determine P_H, P_L, and P_{DC} using SPICE. Plot P_{DC} vs. V_{EE}.
5. Estimate the propagation delays using SPICE and plot the average propagation delay as a function of V_{EE}.

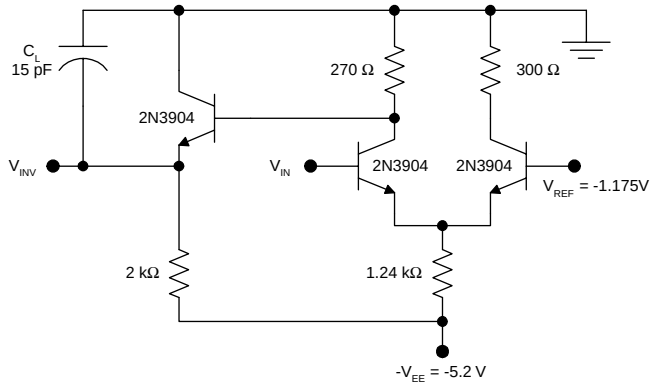


FIGURE 6.43
ECL circuit (L6.4).

6. Build the inverter and measure the voltage transfer characteristic using the x–y feature of an oscilloscope or virtual instrument and a low frequency input signal (1 kHz). Plot the VTCs with V_{EE} as a parameter.

L6.4. A purely capacitive load can slow an ECL circuit considerably if it causes the emitter follower to cut off during a high-to-low transition. Consider the ECL circuit of Figure 6.43 with a lumped capacitive load.

1. Using SPICE, for the inverting output, determine and plot the propagation delays as functions of the load capacitance. Are t_{PLH} and t_{PHL} affected equally by the load capacitance? Why or why not?
2. Build a three-stage ring oscillator on a printed circuit board. Measure the ring oscillator frequency without and with 15-pF loads soldered to each of the three stages. Are the experimental results in agreement with the SPICE predictions?

L6.5. The use of level-sensitive active pull-down circuitry greatly improves the speed of an ECL gate. Consider the LS-APD ECL circuit with a lumped capacitive load as shown in Figure 6.44.

1. Using SPICE, for the inverting output, determine and plot the propagation delays as functions of the load capacitance. Are t_{PLH} and t_{PHL} affected equally by the load capacitance? Why or why not?
2. Build a three-stage ring oscillator on a printed circuit board. Measure the ring oscillator frequency without and with 15-pF loads soldered to each of the three stages. Are the experimental results in agreement with the SPICE predictions?

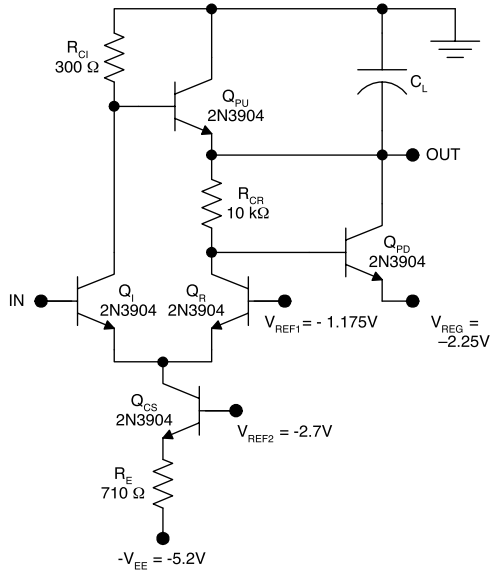


FIGURE 6.44
LS-APD ECL circuit with a lumped capacitive load (L6.5).

Problems

P6.1. Consider the ECL inverter–buffer gate of Figure 6.45. $V_{BEA} = 0.75$ V and $\beta_F = 70$.

1. Determine V_{IL}, V_{IH}, V_{OL}, and V_{OH}.
2. Determine the noise margins V_{NML} and V_{NMH}.
3. Determine the average DC dissipation.

P6.2. For the ECL inverter–buffer shown in Figure 6.46, determine the maximum fan-out assuming that a 0.1-V degradation of V_{OH} is tolerable. $V_{BEA} = 0.75$ V and $\beta_F = 70$.

P6.3. Consider the *unloaded* ECL gate circuit depicted in Figure 6.47.

1. Hand calculate both voltage transfer characteristics.
2. Use SPICE to determine both voltage transfer characteristics. Plot the SPICE and hand-calculated results together for comparison.

P6.4. Consider the *unloaded* ECL circuit illustrated in Figure 6.48.

1. Estimate t_P using approximate hand calculations.
2. Use SPICE to plot the transient response and determine the propagation delays. For the input, use a square wave of appropriate frequency.

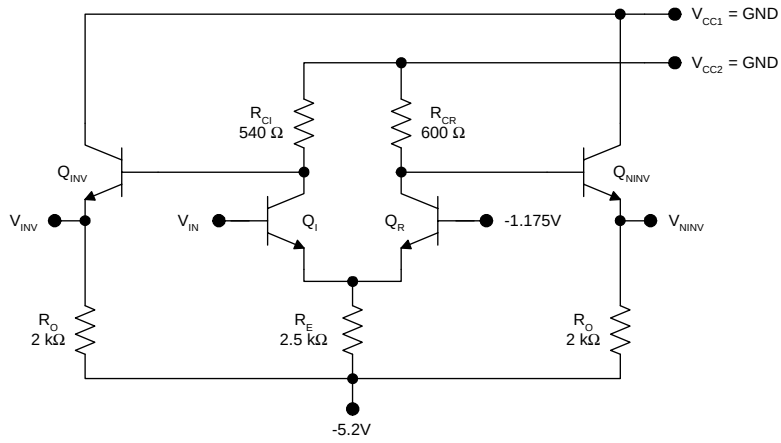


FIGURE 6.45
ECL inverter–buffer gate (P6.1).

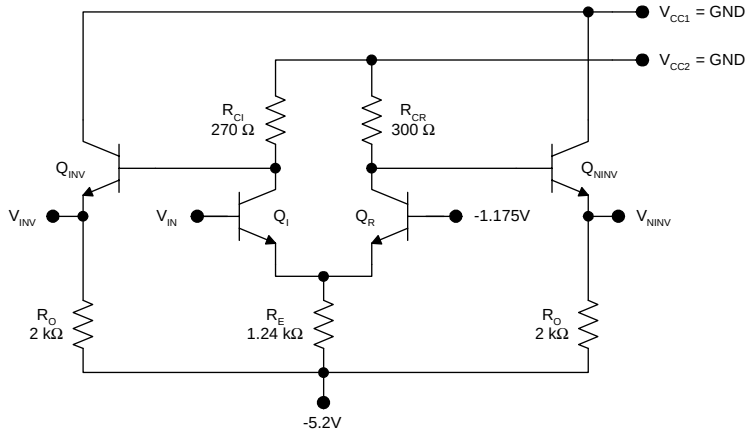


FIGURE 6.46
ECL inverter–buffer (P6.2).

P6.5. Consider the ECL II bias driver shown in Figure 6.49.

1. Estimate the nominal value of V_{REF} at room temperature (300 K).
2. Estimate the temperature derivative for V_{REF} assuming $dV_{BEA}/dT = dV_D/dT = -2$ mV/°C.

P6.6. Using a positive supply with ECL circuitry results in positive ECL (PECL). Consider the version shown in Figure 6.50, which runs off a TTL power supply.

1. Estimate the nominal value of V_{REF} at room temperature (300 K).
2. Calculate and plot the VTCs for the PECL gate.

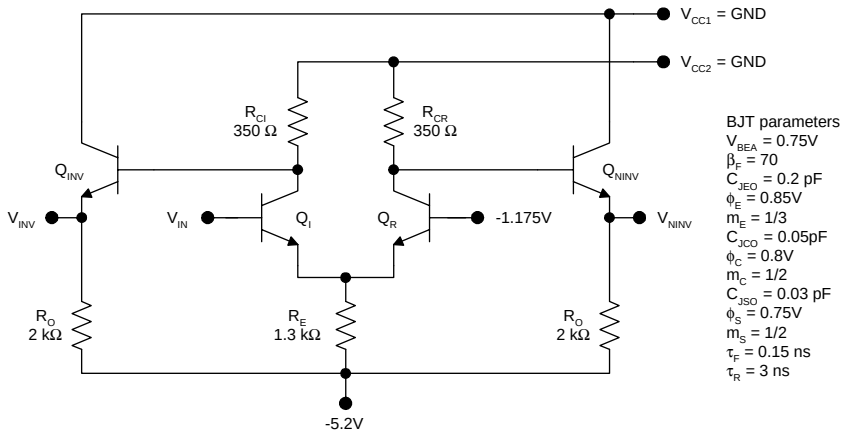


FIGURE 6.47
Unloaded ECL gate circuit (P6.3).

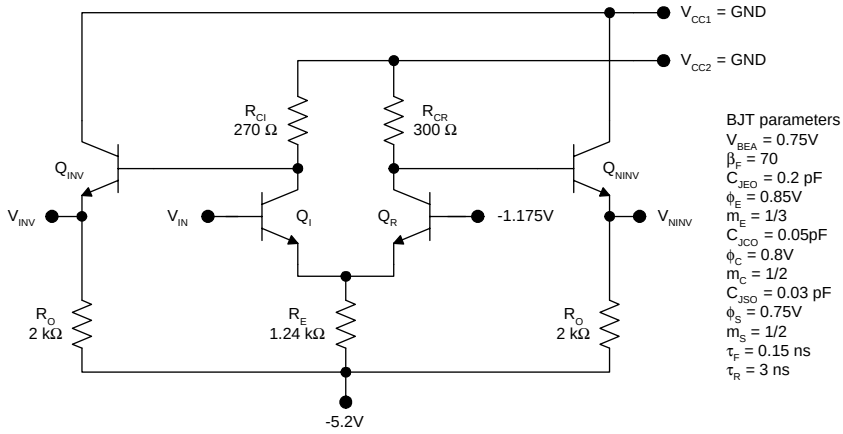


FIGURE 6.48
Unloaded ECL circuit (P6.4).

P6.7. Consider the LS-APD ECL circuit shown in Figure 6.51.
1. Using SPICE, determine the VTC.
2. Using hand analysis, determine P_H and P_L.

P6.8. For the LS-APD ECL gate shown in Figure 6.52, determine and plot t_{PLH} and t_{PHL} vs. the load capacitance.

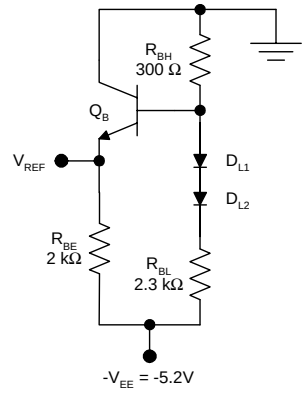


FIGURE 6.49
ECL II bias driver (P6.5).

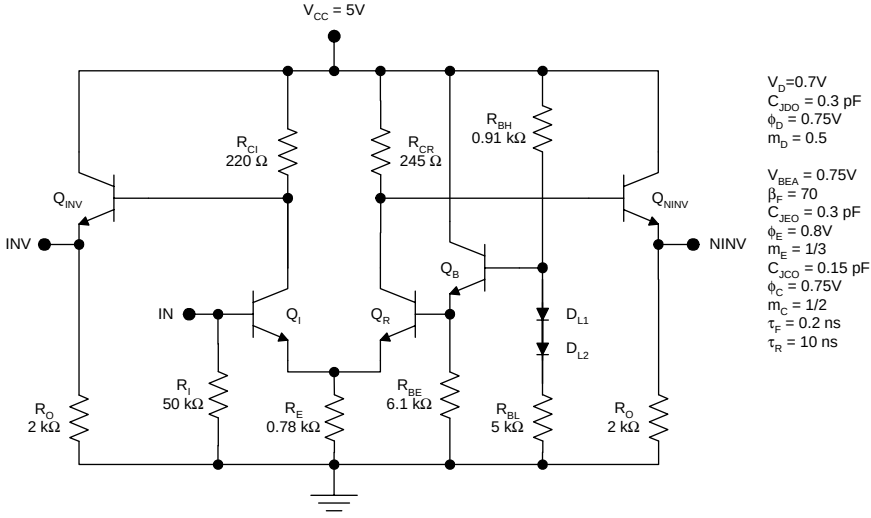


FIGURE 6.50
Positive ECL (PECL) (P6.6).

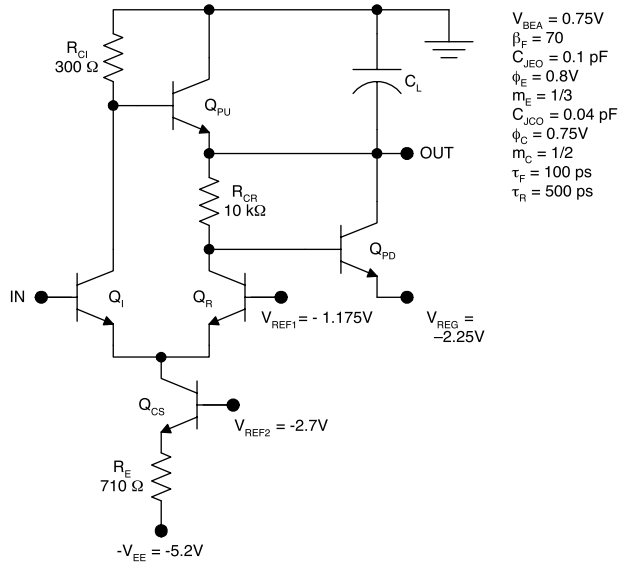


FIGURE 6.51
LS-APD ECL circuit (P6.7).

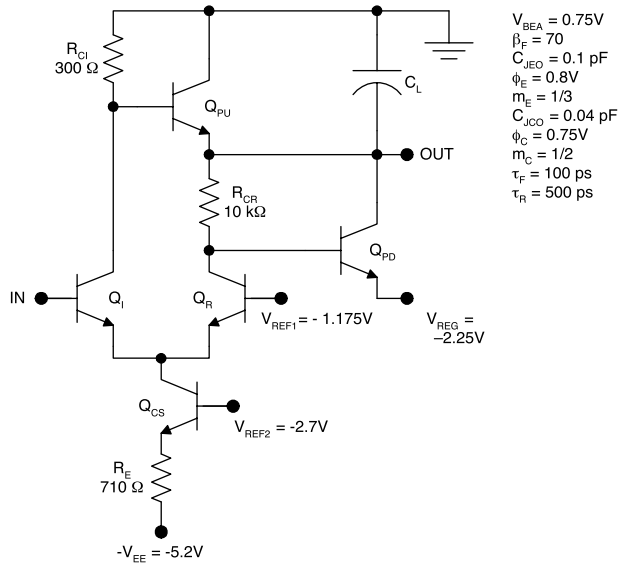


FIGURE 6.52
LS-APD ECL gate (P6.8).

References

1. Toh, K.-Y., Chuang, C.-T., Chen, T.-C., Warnock, J., Li, G.-P., Chin, K., and Ning, T., A 23-ps/2.1-mW ECL gate, *Dig. Tech. Papers 36th IEEE Int. Solid-State Circuits Conf.*, 224, 1989.
2. Uchino, T., Shiba, T., Kikuchi, T., Tamaki, Y., Watanabe, A., Kiyota, Y., and Honda, M., 15-ps ECL/74-GHz f_T Si bipolar technology, *Tech. Dig. 1993 Int. Electron. Devices Meet.*, 67, 1993.
3. Brown, E.W., Agrawal, A., Creary, T., Klein, M.F., Murata, D., and Petolino, J., Implementing Sparc in ECL, *IEEE Micro*, 10, 10, 1990.
4. Wilson, G., Future high performance ECL microprocessors, *Dig. Tech Papers, 1990 Symp. VLSI Circuits*, 5, 1990.
5. Foster, B., Alexander, C., Roberts, A, and Roberts, D., An ECL RISC multiprocessor, *Dig. Papers Compcon Spring '91*, 289, 1991.
6. Gray, D. et al., A 51-K gate low-power ECL gate array family with metal-compiled and embedded SRAM, *Proc. IEEE CICC*, 23.4.1, 1994.
7. Tamamura, M., Shiotsu, S., Hojo, M., Nomura, K., Ichikawa, H., and Akai, T., A 9.5 Gb/s Si bipolar ECL array, *Dig. Tech. Papers 39th Int. Solid-State Circuits Conf.*, 54, 1992.
8. Burghartz, J.N., Comfort, J.H., Patton, G.L., Cressler, J.D., Meyerson, B.S., Stork, J.M.C., Sun, J.Y.-C., Scilla, G., Warnock, J., Ginsberg, B.J., Jenkins, K.A., Toh, K.-Y., Harame, D.L., and Mader, S.R., Sub-30 ps ECL circuits using high-f_T Si and SiGe epitaxial base SEEW transistors, *Tech. Dig. 1990 Int. Electron. Devices Meet.*, 297, 1990.
9. Ghannam, M.Y., Analytical optimization of Si and Si-Ge epitaxial base transistors for very high speed ECL gates, *Proc. 1991 Bipolar Circuits Technol. Meet.*, 166, 1991.
10. Cressler, J.D., Comfort, J.H., Crabbe, E.F., Sun, J.Y.-C., and Stork, J.M.C., An epitaxial emitter cap, SiGe-base bipolar technology with 22 ps ECL gate delay at liquid nitrogen temperature, *Dig. Tech Papers 1992 Symp. VLSI Technol.*, 102, 1992.
11. Shafi, Z.A., Ashburn, P., and Parker, G.J., Predicted propagation delay of Si/SiGe heterojunction bipolar ECL circuits, *IEEE J. Solid-State Circuits*, 25, 1268, 1990.
12. Makowitz, R. and Wild, A., Accurate delay models for ECL logic synthesis, *Proc. 3rd Eur. Conf. Design Automation*, 97 (March 16–19, 1992).
13. O'Brien, P.R., Wyatt, J.L., Jr., Savarino, T.L., and Pierce, J.M., Fast on-chip delay estimation for cell-based emitter coupled logic, *IEEE Int. Symp. Circuits Syst*, 1357, 1988.
14. Chor, E.-F., Brunnschweiler, A., and Ashburn, P., A propagation-delay expression and its application to the optimization of polysilicon emitter ECL processes, *IEEE J. Solid-State Circuits*, 23, 251, 1988.
15. www.motorola.com (Motorola).
16. Sandborn, P.A., Hashemi, H., and Ziai, K., Switching noise in ECL packaging and interconnect systems, *Proc. 1990 Bipolar Circuits Technol. Meet.*, 148, 1990.
17. http://www.fairchildsemi.com (Fairchild Semiconductor).
18. http://www.ti.com (Texas Instruments).

19. Idei, Y., Homma, N., Onai, T., Washio, K., Nishida, T., Nambu, H., and Kanetani, K., Capacitor-coupled complementary emitter-follower for ultra-high-speed low-power bipolar logic circuits, *Dig. Papers Int. Symp. VLSI Circuits*, 25, 1993.
20. Ueda, K., Sasaki, N., Sato, H., and Mashiko, K., A fully compensated active pull-down ECL circuit with self-adjusting driving capability, *IEEE J. Solid State Circ.*, 31, 46, 1996.
21. Kuroda, T., Fujita, T., Noda, M., Thai, P., Yang, L., and Gray, D., Capacitor-free level-sensitive active pull-down ECL circuit with self-adjusting driving capability, *IEEE J. Solid-State Circuits*, 29, 1993.
22. Chuang, C.T., Wu, B., and Anderson, C.J., High-speed low-power cross-coupled active-pull-down ECL circuit, *Proc. 1993 Bipolar/BiCMOS Circuits Technol. Meet.*, 228, 1993.
23. Chuang, C.T., Cressler, J.D., and Warnock, J.D., AC-coupled complementary push-pull ECL circuit with 34-fJ power-delay product, *Electron. Lett.*, 29, 1938, 1993.
24. Chuang, C.T. and Tang, D.D., High-speed low-power AC-coupled complementary push–pull ECL circuit, *IEEE J. Solid-State Circuits*, 27, 660, 1992.
25. Shin, N.J., Lu, P.F., and Chuang, C.T., A high-speed low-power JFET pull-down ECL circuit, *Proc. 1990 Bipolar Circuits Technol. Meet.*, 136, 1990.
26. Chuang, C.T., Chin, K., Shin, H.J., and Lu, P.F, High-speed low-power ECL circuits with AC-coupled self-biased dynamic current source and active-pull-down emitter-follower stage, *IEEE J. Solid-State Circuits*, 27, 1207, 1992.
27. Petty, C. and Pearson, T., Designing with positive ECL, Motorola semiconductor technical data AN1406, http://www.motorola.com, 2001.
28. Operating ECL from a single positive supply, Fairchild Semiconductor application note AN-780, http://www.fairchildsemi.com, 2000.
29. Razavi, B., Ota, Y., and Swartz, R.G., Design techniques for low-voltage high-speed digital bipolar circuits, *IEEE J. Solid-State Circuits*, 29, 332, 1994.
30. Wilhelm, W. and Weger, P., 2-V low-power bipolar logic, *Dig. Tech. Paper. Int. Solid State Circ. Conf.*, 94, 1994.
31. Wisetphanichkij, S., Dejhan, K., Cheevasuvit, F., and Sonyeekan, C., High-speed and low-power ECL circuits design based on BiCMOS technology, *1998 IEEE Asia–Pacific Conf. Circuits Syst.*, 41, 1998.
32. Kuroda, T., Fujita, T., Itabashi, Y., Kabumoto, S., Noda, M., and Kanuma, A., 1.65 Gb/s 60 mW 4:1 multiplexer and 1.8 Gb/s 80 mW 1:4 demultiplexer ICs using 2-V, 3-level series-gating ECL circuits, *Dig. Tech. Papers 42nd Int. Solid-State Circuits Conf.*, 36, 1995.
33. www.cadence.com (Cadence).

7

Field-Effect Transistors

7.1 Introduction

Field-effect transistors (FETs) have several significant differences compared to bipolar junction transistors. First, they are voltage controlled rather than current controlled; this results in low levels of standby supply current and standby power dissipation. Second, they are majority carrier devices — the absence of minority carrier storage effects makes them inherently fast. Third, they can be made smaller than bipolar junction transistors using the same fabrication technology. As a consequence, FET-based logic gates enjoy advantages in performance and packing density over their bipolar counterparts.

Three basic types of FETs are 1) metal oxide–semiconductor field-effect transistor (MOSFET), 2) junction field-effect transistor (JFET), and 3) metal–semiconductor field-effect transistor (MESFET). The MOSFET is most important for digital integrated circuits and will be emphasized in this chapter.

7.1.1 Metal Oxide–Semiconductor Field-Effect Transistor (MOSFET)

The MOSFET, also known as the insulated gate field-effect transistor (IGFET), is the most important device for digital integrated circuits today. A MOSFET device is shown schematically with its circuit symbol in Figure 7.1. The three terminals of this device are the source, gate, and drain, labeled S, G, and D, respectively. Sometimes, a fourth terminal is used: the body or substrate terminal (labeled B). In this device, the voltage applied between the gate and source controls the current flowing between the drain and source.

The basic operation of the MOSFET is as follows. If the gate is biased positively with respect to the source, negatively charged electrons are attracted to the interface between the semiconductor and oxide. This forms a conducting channel between the drain and the source. Then, if the drain is biased positively with respect to the source, electrons in the channel will drift from the source to the drain, resulting in a conventional current flow from drain to source.

In the device of Figure 7.1, the current flow involves only electrons, so it is called an "n-channel" MOSFET. There are also p-channel devices in which

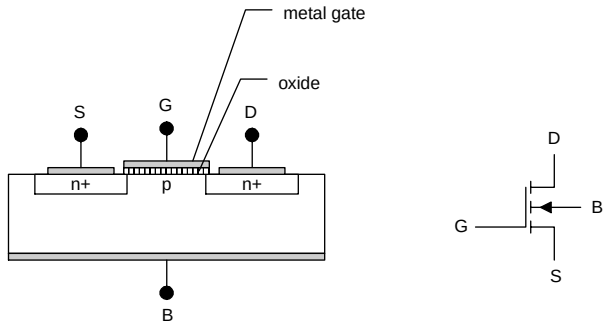


FIGURE 7.1
Enhancement-type n-channel metal oxide–semiconductor field-effect transistor (MOSFET) and circuit symbol.

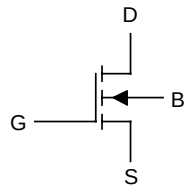
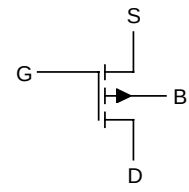
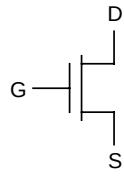
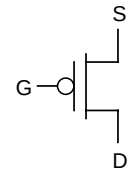
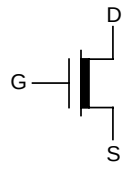
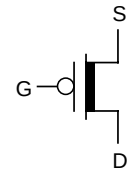
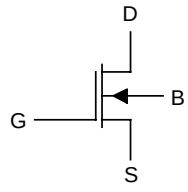
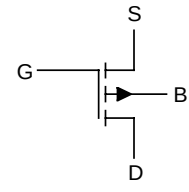
the source and drain are p-type regions and holes drift in the channel. The voltages and currents in the p-channel device have the opposite polarities compared to those in the n-channel device. For the device shown, no conducting channel can be between the drain and source unless a positive voltage is applied between the gate and source; therefore this device is normally off. These MOSFETs are called "enhancement type" because a gate bias is required to enhance a conducting channel. Depletion-type devices are normally on and a gate-to-source bias is necessary to deplete the conducting channel. However, normally off enhancement-type MOSFETs are preferred in digital integrated circuits for low standby dissipation.

Some of the various MOSFET symbols used in the literature are shown in Table 7.1. Of the symbols shown in the table, the most convenient are the middle four. These result in the simplest and neatest circuit diagrams because they eliminate the body connection and avoid the use of other arrows. The inversion circle on the gate indicates a p-type device and the broad line in the channel indicates a depletion-type (normally on) device. These simplified symbols will be used throughout this text, except in situations for which the body bias is used.

7.1.2 Junction Field-Effect Transistor (JFET)

The junction field-effect transistor (JFET)[1,2] takes it name from the gate structure, which involves a p–n junction. Figure 7.2 shows an n-channel device, for which the source, drain, and channel regions are n-type. With zero bias between the gate and source, there is a conducting channel from drain to source; thus, the JFET is a depletion-type device. If a reverse bias is applied to the gate–source junction, this will widen the depletion region and reduce the channel conductivity. A sufficiently negative bias on the gate will pinch off the channel entirely. Therefore, the JFET is a field-effect device in which the gate-to-source bias controls the drain-to-source current. Unlike the case of the MOSFET, no insulating oxide layer is under the gate, so the gate p–n junction must be kept reverse biased in order to avoid a DC gate current.

TABLE 7.1
MOSFET Symbols

	n-channel MOSFETs	p-channel MOSFETs
Enhancement type		
		
Depletion type		
		

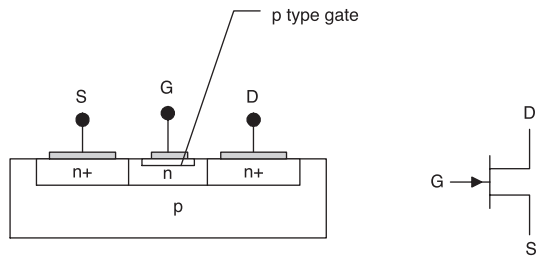


FIGURE 7.2
N-channel junction field-effect transistor (JFET) and circuit symbol.

Although similar, p-channel JFETs utilize p-type regions for the source, drain, and channel. The gate region of a p-channel JFET is doped n-type. For a p-channel JFET, the voltages and currents are reversed in polarity compared to those of an n-channel device. Normally off (enhancement-type)

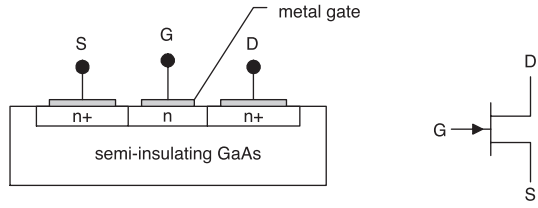


FIGURE 7.3
N-channel gallium arsenide metal–semiconductor field-effect transistor (MESFET) and circuit symbol.

JFETs can be fabricated but with some difficulty. These devices must be made so that the depletion region of the gate junction pinches off the channel at zero gate–source bias. This can be done, but only with precise control of the channel thickness and doping. However, these stringent requirements translate into reduced circuit yield.

JFETs represent important components in analog and mixed-signal integrated circuits but they are not used in digital integrated circuits for two reasons. First, as described earlier, JFETs are inherently depletion-type devices, which results in excessive standby dissipation unless normally off (enhancement-type) devices are fabricated with the concomitant loss of circuit yield. Second, even if normally off JFETs are used, the p–n junctions used in the gates are leaky compared to the MOS structures used in MOSFETs.

7.1.3 Metal–Semiconductor Field-Effect Transistor (MESFET)

The metal oxide–semiconductor field-effect transistor (MESFET)[1,4–6] is similar to the JFET except that a metal–semiconductor junction is used for the gate structure (see Figure 7.3). It suffers from the same drawbacks as the JFET and is not used in silicon technology. MESFETs are used in digital integrated circuits based on compound semiconductors, e.g., gallium arsenide direct-coupled FET logic (DCFL) circuits, because a viable MOSFET technology does not exist in materials such as gallium arsenide and indium phosphide. However, these semiconductors exhibit speed advantages over silicon and are used in some high-end applications.

7.2 MOS Capacitor

MOSFET operation requires that a channel of conducting carriers be controlled by the application of an electric field. The conducting channel exists in the semiconductor while the electric field is established in an oxide insulator layer by a bias on a metal gate. Modulation of the gate bias voltage allows the control of the channel conductivity and therefore the drain current in the MOSFET. This behavior may be understood by considering an MOS capacitor.[1,4–6]

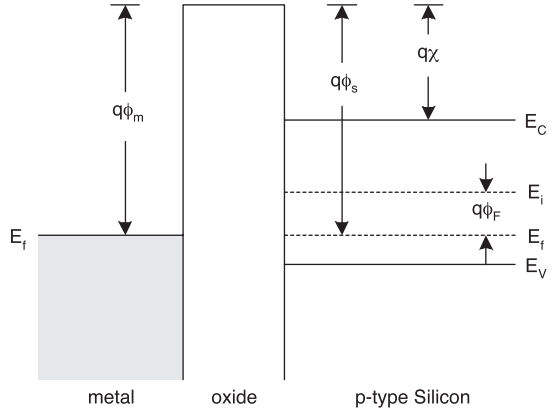


FIGURE 7.4
MOS capacitor under the flat band condition.

Consider a metal oxide–silicon capacitor made on p-type silicon as shown in Figure 7.4. With zero bias applied, the Fermi levels in the metal and semiconductor line up. The difference between the Fermi level in the metal and the vacuum level is the metal work function $q\phi_m$. (This is the amount of energy necessary to remove an electron from the metal to a vacuum.) The difference between the semiconductor conduction band and the vacuum level is the semiconductor electron affinity $q\chi$. The work function for the semiconductor, $q\phi_s$, is a function of the semiconductor electron affinity and the doping in the semiconductor. The separation between the intrinsic Fermi level E_i and the Fermi level E_f in the semiconductor is $q\phi_F$.

If the work functions of the metal and semiconductor are equal and if there are no net charges in the oxide, then zero-applied bias results in the flat band condition depicted in Figure 7.4. With flat bands in the semiconductor, the hole concentration is the same throughout the depth of the p-type silicon.

Applying a positive bias on the metal gate with respect to the semiconductor will result in band bending in the semiconductor to the extent of $q\psi_s$ (Figure 7.5). The offset in the Fermi levels between the semiconductor and the metal is equal to qV, where V is the applied metal–semiconductor bias. Near the silicon–oxide interface, the separation between the Fermi level and the valence band is increased compared to the bulk of the p-type silicon. The hole concentration is therefore reduced near the interface; this situation is referred to as the *depletion* condition.

The application of a sufficiently positive bias on the gate will result in inversion. In this case, the band bending is such that the semiconductor becomes n-type near the interface. It is possible for the semiconductor to be inverted to the extent that the electron concentration near the interface is equal to the hole concentration in the bulk of the semiconductor. This is referred to as *strong inversion* and occurs when the band bending in the semiconductor is given by

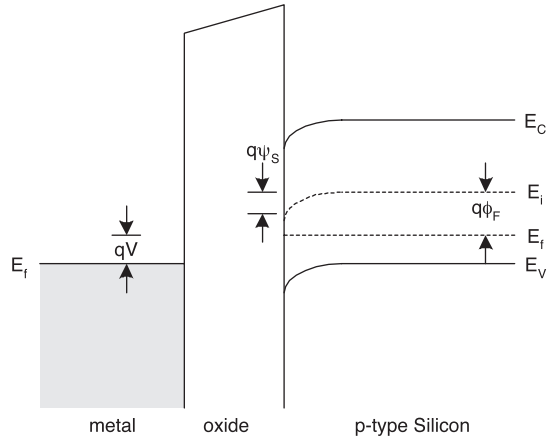


FIGURE 7.5
MOS capacitor under the depletion condition.

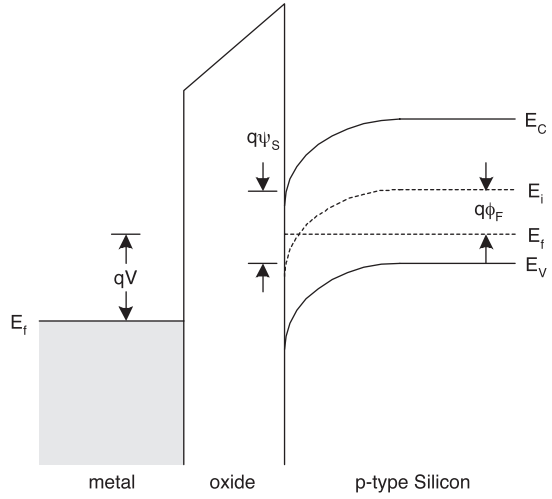


FIGURE 7.6
MOS capacitor under the strong inversion condition.

$$q\psi_S = 2q\phi_F . \tag{7.1}$$

This condition is shown in Figure 7.6.

In an ideal MOS capacitor, the voltage bias that results in strong inversion is

$$V_{inv} = 2\phi_F - \frac{Q_B}{C_{ox}}, \tag{7.2}$$

where $q\phi_F$ is the difference between the intrinsic Fermi level and the Fermi level in the bulk of the p-type semiconductor, C_{OX} is the oxide capacitance, and Q_B is the charge in the depletion layer of the semiconductor under strong inversion. On a per-unit-area basis, the depletion layer charge is

$$\frac{Q_B}{A} = -\sqrt{2\varepsilon_{Si} q N_a |2\phi_F|} \tag{7.3}$$

Equation 7.1 applies to the ideal MOS capacitor, in which the work function difference between the metal and the semiconductor is zero and there is no charge in the oxide.

7.3 MOSFET Threshold Voltage

In an n-channel MOSFET, the gate-to-source bias necessary to cause strong inversion in the channel is called the threshold voltage. Accounting for the difference in the work functions between the metal and semiconductor, and the oxide charge, the threshold voltage is

$$V_{TO} = \phi_{MS} - 2\phi_F - \frac{Q_B}{C_{ox}} - \frac{Q_{ox}}{C_{ox}} - \frac{Q_{II}}{C_{ox}}, \tag{7.4}$$

where

ϕ_{MS} = work function difference between metal and semiconductor
$2\phi_F$ = voltage across the semiconductor necessary to create a conducting channel (inversion layer)
Q_B = charge in the semiconductor under inversion
Q_{OX} = charge in oxide
Q_{II} = charge of ion-implanted impurities in the semiconductor
C_{OX} = oxide capacitance

For an n-channel MOSFET with an acceptor-doped channel region,

$$\phi_F = \frac{kT}{q} \ln\left(\frac{n_i}{N_a}\right). \tag{7.5}$$

If the gate is assumed to be heavily doped polysilicon (*degenerate* polysilicon), then the work function difference is

$$\phi_{MS} = \frac{kT}{q} \ln\left(\frac{n_i}{N_a}\right) - \frac{E_g}{2q}, \tag{7.6}$$

where E_g is the energy gap in silicon (1.12 eV at 300 K). The contribution from the depletion layer charge can be found from

$$\frac{Q_B}{C_{ox}} = \frac{Q_B/A}{C_{ox}/A} = \frac{-\sqrt{2qN_a\varepsilon_{Si}|2\phi_F|}}{\varepsilon_{ox}/t_{ox}}. \tag{7.7}$$

A nonzero bias, V_{BS}, applied between the body and the source of the MOSFET further modifies the threshold voltage to

$$V_T = V_{TO} + \gamma\left(\sqrt{|V_{BS} + 2\phi_F|} - \sqrt{|2\phi_F|}\right), \tag{7.8}$$

where γ is the body effect coefficient given by

$$\gamma = \frac{\sqrt{2q\varepsilon_{Si}N_a}}{C_{ox}/A}, \tag{7.9}$$

where

q = electronic charge, 1.602×10^{-19}C
ε_{Si} = permittivity of silicon
N_a = acceptor doping in the p-type silicon (n-channel MOSFET)
C_{OX}/A = oxide capacitance per unit area

The most significant aspect of the threshold voltage is that it can be controlled *precisely* by the ion implantation of impurities (typically As, P, or B) into the silicon. Using As or P makes the threshold voltage more positive, whereas the use of B makes it more negative; the actual threshold voltage can be either sign. Among n-channel MOSFETs, enhancement-type transistors have positive thresholds and depletion-type transistors have negative thresholds. The opposite is true for p-channel devices.

It is also important to note that the body bias effect allows the threshold of a MOSFET to be adjusted *in the circuit*. This is exploited in active biasing schemes to control subthreshold conduction and to overcome manufacturing tolerances in the threshold voltages. Both of these techniques are utilized in modern low-power, high-speed CMOS circuits.

Example 7.1

Calculate the zero-bias threshold voltage for an n-channel MOSFET with t_{OX} = 7.5 nm and $N_a = 10^{16}$ cm^{-3}. Assume that the gate is heavily doped n-polysilicon (with the Fermi level coincident with the conduction band) and that 10^{11} cm^{-2} positive charges are in the oxide. A boron dose of 10^{12} cm^{-2} is implanted to adjust the threshold voltage.

Solution. The zero-bias threshold is

$$V_{TO} = \phi_{MS} - 2\phi_F - \frac{Q_B}{C_{ox}} - \frac{Q_{ox}}{C_{ox}} - \frac{Q_{II}}{C_{ox}}$$

and, for this level of doping,

$$\phi_F = \frac{kT}{q}\ln\left(\frac{n_i}{N_a}\right) = (0.026\text{ V})\ln\left(\frac{1.45\times10^{10}\text{ cm}^{-3}}{10^{16}\text{ cm}^{-3}}\right) = -0.35\text{ V}.$$

The work function difference is

$$\phi_{MS} = \frac{kT}{q}\ln\left(\frac{n_i}{N_a}\right) - \frac{E_g}{2q} = -0.35\text{ V} - \frac{1.12\text{ V}}{2} = -0.90\text{ V}.$$

The contribution due to the depletion charge in the semiconductor under inversion is

$$\frac{Q_B}{C_{ox}} = \frac{-\sqrt{2qN_a\varepsilon_{Si}|2\phi_F|}}{\varepsilon_{ox}/t_{ox}}$$

$$= \frac{-\sqrt{2\left(1.602\times10^{-19}\text{ C}\right)\left(10^{16}\text{ cm}^{-3}\right)(11.9)\left(8.85\times10^{-14}\text{ F/cm}\right)|2(-0.35\text{ V})|}}{(3.9)\left(8.85\times10^{-14}\text{ F/cm}\right)/7.5\times10^{-7}\text{ cm}}$$

$$= -0.106\text{ V}$$

the contribution due to the oxide charge is

$$\frac{Q_{ox}}{C_{ox}} = \frac{\left(1.602\times10^{-19}\text{ C}\right)\left(10^{11}\text{ cm}^{-2}\right)}{(3.9)\left(8.85\times10^{-14}\text{ F/cm}\right)/7.5\times10^{-7}\text{ cm}} = 0.035\text{ V},$$

and the adjustment due to the implantation of boron is

$$\frac{Q_{II}}{C_{ox}} = -\frac{\left(1.602\times10^{-19}\text{ C}\right)\left(10^{12}\text{ cm}^{-2}\right)}{(3.9)\left(8.85\times10^{-14}\text{ F/cm}\right)/7.5\times10^{-7}\text{ cm}} = -0.35\text{ V}.$$

Therefore, the zero-bias threshold voltage is

$$V_{TO} = \phi_{MS} - 2\phi_F - \frac{Q_B}{C_{ox}} - \frac{Q_{ox}}{C_{ox}} - \frac{Q_{II}}{C_{ox}}$$

$$= -0.90\text{ V} + 0.70\text{ V} + 0.106\text{ V} - 0.035\text{ V} + 0.35\text{ V}$$

$$= 0.22\text{ V}$$

The ion implantation adjustment was necessary to obtain an enhancement-type transistor with a positive threshold voltage.

Example 7.2

Consider an n-channel MOSFET with t_{OX} = 7.5 nm and N_a = 10^{16} cm^{-3}. Calculate the change in the threshold voltage due to the body bias effect if V_{BS} = –1.5 V.

Solution. The body effect coefficient is

$$\gamma = \frac{\sqrt{2q\varepsilon_{Si}N_a}}{C_{ox}/A}$$

$$= \frac{\sqrt{2\left(1.602\times10^{-19}\text{ C}\right)(11.9)\left(8.85\times10^{-14}\text{ F/cm}\right)\left(10^{16}\text{ cm}^{-3}\right)}}{(3.9)\left(8.85\times10^{-14}\text{ F/cm}\right)/7.5\times10^{-7}\text{ cm}}$$

$$= 0.126\text{ V}^{1/2}$$

Also,

$$\phi_F = \frac{kT}{q}\ln\left(\frac{N_a}{n_i}\right) = 0.026\text{ V}\ln\left(\frac{10^{16}\text{ cm}^{-3}}{1.45\times10^{10}\text{ cm}^{-3}}\right) = 0.35\text{ V};$$

therefore,

$$\Delta V_T = V_T - V_{TO}$$

$$= \gamma\left(\sqrt{\left|V_{BS}+2\phi_F\right|} - \sqrt{\left|2\phi_F\right|}\right)$$

$$= \left(0.126\text{ V}^{1/2}\right)\left(\sqrt{2.20\text{ V}} - \sqrt{0.70\text{ V}}\right)$$

$$= 0.08\text{ V}$$

Here, a negative bias is on the body with respect to the source. This repels electrons from the channel and makes the threshold voltage more positive.

7.4 Long-Channel MOSFET Operation

Consider an n-channel enhancement-type MOSFET biased as in Figure 7.7. Here the substrate connection has not been shown explicitly; assume that the substrate is shorted to the source (as is often the case). V_{GS} is the gate-to-source bias, V_{DS} is the drain-to-source bias, and I_D is the drain current.

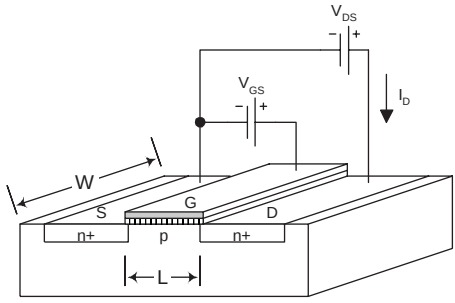


FIGURE 7.7
N-channel enhancement-type MOSFET with bias.

The MOSFET has three modes of operation: *cutoff*, *linear*, and *saturation*. *Cutoff* occurs if the gate-to-source bias voltage is insufficiently positive to induce a conducting channel. As a first-order approximation, cutoff results in zero drain current. If the gate-to-source bias is made more positive than the threshold voltage for the device, then a conducting channel is induced and a drain current can flow. With a small drain-to-source bias, the MOSFET acts like a voltage-controlled resistance. This is the *linear* (also known as "ohmic" or "triode") mode of operation. However, if the drain-to-source bias is sufficiently large, then the conducting channel will pinch off at the drain end. This causes the drain current to saturate, so this mode of operation is called *saturation*.

The characteristic curves for a MOSFET are shown in Figure 7.8. In such characteristics it is customary to plot the drain current, I_D, vs. the drain-to-source voltage, V_{DS}, with the gate-to-source voltage, V_{GS}, as a parameter. This results in a family of curves, one for each particular value of V_{GS}. Cutoff is associated with zero drain current, so its locus is on the V_{DS} axis. In the linear region, the drain current increases approximately linearly with the drain-to-source voltage; its locus is to the left of the parabola. Saturation is characterized by a constant drain current and its locus is to the right of the parabola. Each of these modes of operation will be discussed in more detail in the following sections.

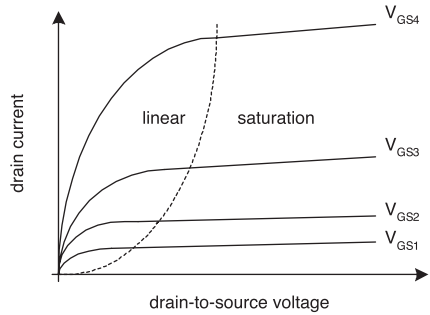


FIGURE 7.8
Characteristic curves for a MOSFET.

7.4.1 MOSFET Cutoff Operation

Cutoff operation occurs if the gate-to-source bias is less positive than the threshold voltage. To a first approximation, cutoff is accompanied with zero drain current:

$$I_D \approx 0 \quad \left(V_{GS} < V_T\right). \tag{7.10}$$

7.4.2 MOSFET Linear Operation

Linear operation occurs if the gate-to-source bias is more positive than the threshold and the drain-to-source bias is small enough so that the channel does not pinch off at the drain end. In the linear mode of operation, the MOSFET acts like a voltage-controlled resistance. The controlling variable is the gate-to-source voltage and the controlled variable is the drain-to-source resistance. The drain current in the linear mode of operation may be determined with the aid of Figure 7.9.

It is customary to assume that carriers move only by drift and that the drift is only in the direction parallel to the interface (the *gradual channel approximation*).[1,4] Based on this assumption, the drain current at a point, *y*, along the channel is

$$I_D(y) = qW \int_{x=0}^{\infty} \mu_n n(x,y) \frac{dV}{dy} dx. \tag{7.11}$$

Integration over *x* (the depth of the inversion layer of conducting electrons) yields

$$I_D(y) = -\mu_n W \frac{dV}{dy} Q_i(y), \tag{7.12}$$

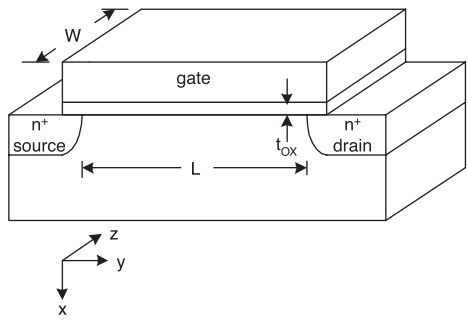


FIGURE 7.9
MOSFET for determination of the drain current.

where $Q_i(y)$ is the integrated electron charge in the inversion layer per unit area at a distance along the channel, y. If the inversion layer charge is assumed to be located in a sheet of zero thickness (the *charge sheet approximation*)[7] and the bulk depletion charge to be approximately independent of the gate-to-source bias, then

$$Q_i(y) \approx -\frac{\varepsilon_{ox}}{t_{ox}}\left(V_{GS} - V(y) - V_T\right). \tag{7.13}$$

Therefore,

$$I_D dy = \frac{\mu_n \varepsilon_{ox}}{t_{ox}} W\left(V(y) - V_T\right) dV . \tag{7.14}$$

Integrating over the length of the channel yields

$$\int_0^L I_D dy = \frac{\mu_n \varepsilon_{ox}}{t_{ox}} W \int_0^{V_{DS}} \left(V_{GS} - V(y) - V_T\right) dV , \tag{7.15}$$

which results in the well-known equation for drain current in the linear region of operation:

$$I_D = \frac{\mu_n \varepsilon_{ox} W}{t_{ox} L}\left[\left(V_{GS} - V_T\right)V_{DS} - \frac{V_{DS}^2}{2}\right] . \tag{7.16}$$

This equation is subject to several limitations. First, it was developed based on the assumption that carriers move by drift only and at the low-field mobility. However, carriers move by diffusion as well as drift near the threshold. Second, in short-channel transistors, it is not valid to assume that carriers drift according to the low-field mobility over the entire channel length; carrier velocity saturation must be considered. Third, the variation of the depletion charge with the gate-to-source bias has been neglected in comparison to the inversion charge. This approximation is reasonable for lightly doped channels.

The transistor operates in the linear region as long as the gate-to-source bias is greater than the threshold voltage and the channel does not pinch off at the drain end. Pinch-off at the drain end of the channel occurs when

$$V_{GS} - V_{DS} = V_T . \tag{7.17}$$

This condition defines the boundary between linear and saturation operation.

Usually the linear drain current equation is written in terms of the *device transconductance parameter*:

$$I_D = K\left[\left(V_{GS} - V_T\right)V_{DS} - \frac{V_{DS}^2}{2}\right], \qquad V_{GS} \geq V_T \text{ and } \left(V_{GS} - V_T\right) \geq V_{DS}, \quad (7.18)$$

where K is the device transconductance parameter given by

$$K = \frac{W}{L}\frac{\mu_n \varepsilon_{ox}}{t_{ox}}, \quad (7.19)$$

where

W = width of device
L = length of device
μ_n = mobility of electrons (n-channel device)
ε_{OX} = permittivity of oxide
t_{OX} = oxide thickness

Sometimes the device transconductance parameter is calculated as

$$K = \frac{W}{L}k', \quad (7.20)$$

where k' is the *process transconductance parameter* given by

$$k' = \frac{\mu_n \varepsilon_{ox}}{t_{ox}}. \quad (7.21)$$

Another useful relationship for linear operation allows calculation of the drain-to-source voltage, if the gate-to-source voltage and the drain current are known. If the channel length modulation is neglected, then using the quadratic formula yields

$$V_{DS} = \left(V_{GS} - V_T\right) - \sqrt{\left(V_{GS} - V_T\right)^2 - \frac{2I_D}{K}}, \quad V_{GS} \geq V_T \text{ and } \left(V_{GS} - V_T\right) \geq V_{DS}. \quad (7.22)$$

The relationships for p-channel MOSFETs may be obtained in similar manner. However, μ_p must be used instead of μ_n in all of the preceding equations and all voltages and currents are opposite in polarity.

Example 7.3

Estimate the process transconductance parameters for n-channel and p-channel MOSFETs with t_{OX} = 10 nm.

Solution. For n-channel MOSFETs, assuming $\mu_n = 580\ \text{cm}^2\text{V}^{-1}\text{s}^{-1}$, the process transconductance parameter is

$$k'_N = \frac{\mu_n \varepsilon_{ox}}{t_{ox}} = \frac{(580\,\text{cm}^2\text{V}^{-1}\text{s}^{-1})(3.9)(8.85\times10^{-14}\,\text{F/cm})}{10\times10^{-7}\,\text{cm}} = 200\ \mu\text{A/V}^2 .$$

For p-channel MOSFETs, assuming $\mu_p = 230\ \text{cm}^2\text{V}^{-1}\text{s}^{-1}$, the process transconductance parameter is

$$k'_P = \frac{\mu_p \varepsilon_{ox}}{t_{ox}} = \frac{(230\ \text{cm}^2\text{V}^{-1}\text{s}^{-1})(3.9)(8.85\times10^{-14}\,\text{F/cm})}{10\times10^{-7}\,\text{cm}} = 80\ \mu\text{A/V}^2 .$$

The process transconductance parameter is typically 2.5 times larger for n-channel MOSFETs than for p-channel MOSFETs.

Example 7.4

Estimate the device transconductance parameters for n-channel and p-channel MOSFETs with $t_{OX} = 10$ nm, $W = 2$ μm, and $L = 0.25$ μm.

Solution. For the n-channel MOSFET, the device transconductance parameter is

$$K_N = \frac{W}{L}k'_N = \left(\frac{2\ \mu\text{m}}{0.25\ \mu\text{m}}\right)200\ \mu\text{A/V}^2 = 1.60\ \text{mA/V}^2 .$$

For the p-channel MOSFET, the device transconductance parameter is

$$K_P = \frac{W}{L}k'_p = \left(\frac{2\ \mu\text{m}}{0.25\ \mu\text{m}}\right)80\ \mu\text{A/V}^2 = 0.64\ \text{mA/V}^2 .$$

If n- and p-channel MOSFETs are fabricated with the same aspect ratio, the n-channel MOSFET has a device transconductance parameter 2.5 times larger than that of the p-channel MOSFET.

7.4.3 MOSFET Saturation Operation

Saturation operation occurs if the gate-to-source bias is more positive than the threshold and the drain-to-source bias is large enough to cause the channel to pinch off at the drain end. Pinch-off at the drain end of the channel occurs when

$$V_{DS} = V_{GS} - V_T . \tag{7.23}$$

Substituting this result in the equation for linear operation yields the equation for saturated operation of the MOSFET:

$$I_D = \frac{K}{2}\left(V_{GS} - V_T\right)^2; \qquad V_{GS} \geq V_T \text{ and } \left(V_{GS} - V_T\right) \leq V_{DS}. \tag{7.24}$$

Thus, in the saturation mode of operation, the MOSFET acts like a voltage-controlled current source. The drain current is the controlled quantity and the gate-to-source bias is the controlling quantity.

7.4.4 MOSFET Subthreshold Operation

Cutoff operation of the MOSFET ($V_{GS} < V_T$ for an n-MOSFET or $|V_{GS}| < |V_T|$ for a p-MOSFET) results in zero drain current, to a first approximation. However, if the gate-to-source bias voltage is close to the threshold voltage, a non-negligible drain current will flow. This *subthreshold* current is important in modern low-voltage, low-power CMOS and memory circuits.

Although saturation or linear operation of the MOSFET is dominated by the drift of majority carriers, subthreshold operation occurs as the result of minority carrier diffusion.[1,4,6] Essentially, the device acts as a bipolar transistor in which the source injects carriers into the channel region. These injected carriers diffuse the length of the channel and are collected by the drain. In an n-MOSFET, for example, electrons are injected into the (noninverted) p-type channel region and diffuse to the drain, resulting in conventional current flow from the drain to the source. In other words, the subthreshold current flows in the same direction as the saturated current.

If it is assumed that diffusion alone is responsible for the subthreshold current, then

$$I_D = -qAD_n \frac{dn}{dy} = qAD_n \frac{n(0) - n(L)}{L}, \tag{7.25}$$

where A is the effective cross section for the current flow. The electron concentrations at the source and drain ends of the channel are approximately

$$n(0) \approx \overline{n}_{po} \exp\left(\frac{q\psi_s}{kT}\right) \tag{7.26}$$

and

$$n(L) \approx \overline{n}_{po} \exp\left(\frac{q\left(\psi_s - V_{DS}\right)}{kT}\right), \tag{7.27}$$

where ψ_s is the band bending in the semiconductor. The effective thickness of the inversion layer is kT/qE_s, where E_s is the surface electric field intensity. Thus

$$A \approx \frac{kTW}{q}\sqrt{\frac{\varepsilon_{Si}}{2qN_a\psi_s}}, \tag{7.28}$$

where the band bending in the semiconductor is given by

$$\psi_s \approx 2\psi_B + (1+m)\left(V_{GS} - V_T\right). \tag{7.29}$$

The unitless parameter m is given by

$$m = 1 + \frac{C_{dm}}{C_{ox}}, \tag{7.30}$$

where

C_{dm} = maximum depletion layer capacitance of the semiconductor under the oxide
C_{OX} = oxide capacitance

Combining the previous equations and making use of the Einstein relationship yields the subthreshold current in the n-MOSFET:

$$I_D = \frac{\mu_n \varepsilon_{ox} W(1+m)}{t_{ox} L}\left(\frac{kT}{q}\right)^2 \exp\left(\frac{q\left(V_{GS} - V_T\right)}{mkT}\right)\left[1 - \exp\left(-\frac{qV_{DS}}{kT}\right)\right]. \tag{7.31}$$

In typical MOSFETs, $1 < m < 2$.*

If the drain-to-source bias is several times $kT/q \sim 26$ mV at room temperature, then the subthreshold current is independent of the drain-to-source bias:

$$I_D \approx K(1+m)\left(\frac{kT}{q}\right)^2 \exp\left(\frac{q\left(V_{GS} - V_T\right)}{mkT}\right). \tag{7.32}$$

An important figure for subthreshold operation is the *subthreshold swing*, defined as

$$S \equiv \left(\frac{d\left(\log_{10} I_D\right)}{dV_{GS}}\right)^{-1}. \tag{7.33}$$

* m was assumed to be unity for the analysis of the linear and saturation regions of operation. This amounts to neglecting the change in the depletion layer charge in comparison to the inversion layer charge; however, this approximation is not appropriate for the analysis of subthreshold operation.

From Equation 7.32, the subthreshold swing is

$$S = 2.3\frac{mkT}{q}. \tag{7.34}$$

Typically, room-temperature operation of MOSFETs is characterized by a subthreshold swing of 100 mV. This means that the subthreshold current changes by one decade for every 100-mV change in the gate-to-source bias. The practical implication of this is that scaling of MOSFET threshold voltages below about $|V_T| < 300$ mV is accompanied by significant subthreshold current at $V_{GS} = 0$. As will be shown in Chapter 10, this is a significant issue in the design of low-power CMOS circuits.

Example 7.5

Calculate and plot the subthreshold drain current for an n-channel MOSFET with $K = 1.6$ mA/V^2, $V_T = 0.3$ V, and $m = 1.6$.

Solution. Assuming $V_{DS} > 3\,kT/q$, the subthreshold current is given by

$$\begin{aligned} I_D &\approx K(1+m)\left(\frac{kT}{q}\right)^2 \exp\left(\frac{q(V_{GS}-V_T)}{mkT}\right) \\ &= (1.6\text{ mA/V}^2)(2.6)(26\text{ mV})^2 \exp\left(\frac{V_{GS}-V_T}{(1.6)(26\text{ mV})}\right) \\ &= (2.8\ \mu\text{A})\exp\left(\frac{V_{GS}-V_T}{41\text{ mV}}\right) \end{aligned}$$

The results are plotted in Figure 7.10. The characteristic is a straight line on a semilog plot, showing that the subthreshold current increases exponentially with the gate-to-source bias.

7.4.5 Transit Time

It takes a finite time for majority carriers to traverse the channel in a conducting MOSFET — a delay called the transit time, t_t. For the purpose of estimating propagation delays in MOS circuits, the transit time is usually assumed to be much shorter than the circuit delays; this is called the quasi-static assumption. However, in MOS circuits with very little external loading, the transit time presents a fundamental limitation to the switching speed.

In a long-channel n-channel MOSFET, the drift of electrons in the channel is governed by Ohm's law and the low-field mobility. The average electric field intensity in the channel is approximately

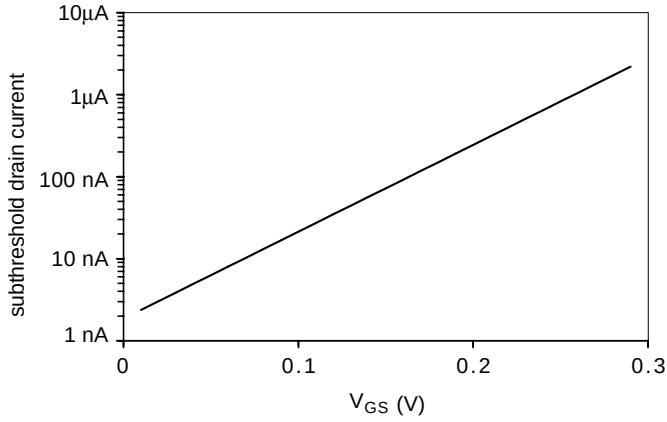


FIGURE 7.10
Subthreshold current in an n-channel MOSFET with K = 1.6 mA/V², V_T = 0.3 V, and m = 1.6.

$$E \approx \frac{V_{DS}}{L}. \tag{7.35}$$

Therefore, the carriers move at a velocity of approximately

$$v \approx \frac{\mu_n V_{DS}}{L}, \tag{7.36}$$

so the transit time is

$$t_t = \frac{L}{v} = \frac{L^2}{\mu_n V_{DS}}. \tag{7.37}$$

The transit time thus increases with the square of the channel length.

Example 7.6

Estimate the transit time for n-channel MOSFETs with a 5-μm channel length in a 5-V CMOS circuit.

Solution. Assuming an electron mobility of 580 cm²/Vs for electrons, the transit time is

$$t_t = \frac{L^2}{\mu_n V_{DS}} = \frac{\left(5\times10^{-4}\,\text{cm}\right)^2}{\left(580\ \text{cm}^2\text{V}^{-1}\text{s}^{-1}\right)(5\ \text{V})} = 86\ \text{ps}.$$

The quasi-static approximation is applicable if the circuit propagation delay is greater than the transit time.

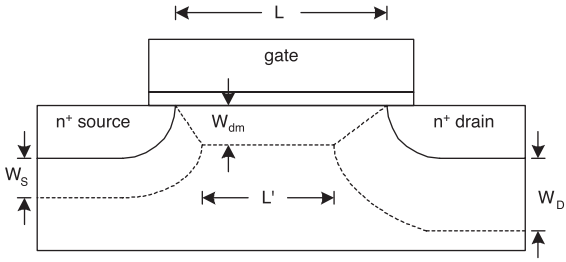


FIGURE 7.11
Charge sharing in a short-channel MOSFET.

7.5 Short-Channel MOSFETs

Aggressive scaling of MOSFETs and, in particular, the channel lengths in MOSFETs has resulted in devices that behave differently than the long-channel devices described in Section 7.4. First, the threshold voltage becomes a function of the channel length (the so-called *short-channel effect*). Second, the electric field intensity in the channel may be sufficiently large so that the carriers reach their saturated velocity. Third, the effective channel length becomes a function of the drain-to-source bias as a consequence of *channel length modulation*. All of these effects are of practical importance in the design of high-performance CMOS circuits today.

7.5.1 The Short-Channel Effect

As a consequence of the short-channel effect (SCE), the absolute value of the threshold voltage decreases with decreasing channel length. This may be understood on the basis of a charge-sharing model, as illustrated in Figure 7.11.[6,8,9]

In the short-channel MOSFET, some of the field lines in the source and drain depletion regions terminate on charges under the gate. In other words, some of the depletion layer charge under the gate is shared with the source and drain. Therefore, the threshold voltage should be estimated based on the trapezoidal region of charge under the gate. Then the threshold voltage is reduced, compared to the long-channel value, by

$$\Delta V_T \approx \frac{\varepsilon_{ox}(L'-L)W_{dm}}{2WLt_{ox}} \approx -\frac{\varepsilon_{ox}(W_S+W_D)W_{dm}}{2WLt_{ox}}, \tag{7.38}$$

where

ε_{OX} = permittivity of oxide
t_{OX} = oxide thickness

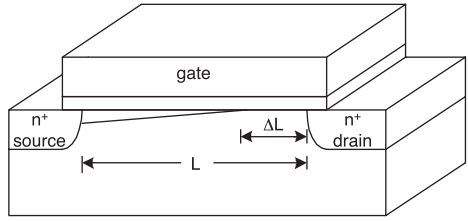


FIGURE 7.12
Channel length modulation in a short-channel MOSFET.

W = width of MOSFET channel
L = length of MOSFET channel
W_{dm} = depletion width in semiconductor under inversion
W_S = source junction depletion width
W_D = drain junction depletion width

In practice, careful design can overcome the short-channel effect because all of the MOSFETs may have the minimum channel length and, therefore, the same threshold voltage. However, in small-width devices the threshold also becomes dependent on the device width because of charge sharing. This requires that the implantation adjustment be designed so that the narrowest devices on the wafer have acceptable threshold voltages.

7.5.2 Channel Length Modulation

The drain current in a MOSFET saturates at the value of V_{DS}, which causes the channel to pinch off at the drain end. Further increase in V_{DS} causes the pinch-off point to move into the channel, toward the source, as shown in Figure 7.12.

This increases the drain current by the ratio $L/(L - \Delta L)$. In a long-channel MOSFET, this effect is inconsequential because the percentage change in the drain current is small; however, the *channel length modulation effect* is important in short-channel MOSFETs.[1,4,6] Mathematically, the channel length modulation is modeled by multiplying the drain current expressions by a factor that increases linearly with the drain-to-source bias. For linear operation, the expression for the drain current including the channel length modulation is

$$I_D = K\left[\left(V_{GS} - V_T\right)V_{DS} - \frac{V_{DS}^2}{2}\right]\left[1 + \lambda V_{DS}\right], \quad V_{GS} \geq V_T \text{ and } \left(V_{GS} - V_T\right) \geq V_{DS}, \tag{7.39}$$

where λ is the empirical *channel length modulation parameter*. Similarly, the equation for the saturated drain current becomes

$$I_D = \frac{K}{2}\left(V_{GS} - V_T\right)^2\left(1 + \lambda V_{DS}\right); \qquad V_{GS} \geq V_T \text{ and } \left(V_{GS} - V_T\right) \leq V_{DS}. \tag{7.40}$$

7.5.3 Velocity Saturation

At high electric-field intensities, the carrier drift velocities are no longer proportional to the electric field. Instead, there is approximately carrier velocity saturation — an important effect in short-channel MOSFETs. In such devices, the onset of drain current saturation occurs at a lower drain-to-source bias than predicted in Section 7.4. Also, the magnitude of the saturated drain current is less than that predicted on the basis of the low-field mobility.[10]

Empirically, the carrier velocity vs. field characteristics can be fit to[11,12]

$$v = \frac{\mu E}{\left[1 + \left(\frac{\mu E}{v_{sat}}\right)^n\right]^{1/n}}, \tag{7.41}$$

where

μ = carrier mobility
E = electric field intensity
v_{sat} = carrier saturation velocity
n = an empirical parameter

For holes in silicon, $n = 1$ and $v_{sat} = 8 \times 10^6$ cm/s. For electrons in silicon, $n = 2$ and $v_{sat} = 10^7$ cm/s. (The saturation velocities in silicon MOSFETs are typically 20% lower than the bulk values.) For the case of n = 1, the linear drain current is decreased to

$$I_D = \frac{K\left[\left(V_{GS} - V_T\right)V_{DS} - \frac{V_{DS}^2}{2}\right]}{1 + \frac{\mu V_{DS}}{v_{sat} L}}. \tag{7.42}$$

Carrier velocity saturation reduces the saturated drain current to

$$I_D = \frac{\varepsilon_{ox} W v_{sat}}{t_{ox}}\left(V_{GS} - V_T\right). \tag{7.43}$$

Therefore, the drain current is reduced compared to the constant mobility case. Also, *the saturated drain current becomes proportional to the device width, not the aspect ratio*. The $n = 2$ case is qualitatively similar, but the analysis of the linear drain current is considerably more complex.

7.5.4 Transit Time in Short-Channel MOSFETs

In short-channel MOSFETs, the carriers may travel at close to the saturation velocity for the entire length of the channel. In this limit, the transit time is

$$t_t = \frac{L}{v_{sat}}; \tag{7.44}$$

therefore, the transit time is directly proportional to the channel length in short-channel MOSFETs.

Example 7.7

Estimate the transit time for n-channel MOSFETs with a 65-nm channel length in a 1.5-V CMOS circuit.

Solution. The average channel field for a saturated MOSFET is approximately

$$E \approx \frac{1.5\text{ V}}{65\times10^{-7}\text{ cm}} = 2.3\times10^{5}\text{ V/cm}.$$

Therefore, it is appropriate to assume that the electrons drift at their saturated velocity in the channel. The transit time is

$$t_t = \frac{L}{v_{sat}} = \frac{65\times10^{-7}\text{ cm}}{10^{7}\text{ cm/s}} = 0.65\text{ps}.$$

The quasi-static approximation is applicable if the circuit propagation delay is greater than this transit time.

7.6 MOSFET SPICE Models

Four MOSFET models are used by SPICE:

- Level 1 is simplest, allowing short computation times while providing reasonable accuracy for most purposes.
- Level 2 is more accurate, accounting for the bias-dependent carrier mobility and carrier velocity saturation and providing a more accurate model for the channel length modulation. However, the complexity of its equations results in longer computation times.
- Level 3 was developed to provide accuracy similar to the level 2 model but with shorter computational times.
- Level 4, which is empirical rather than analytical, provides the best accuracy for short-channel devices with reasonable computational times. There are several versions of the level 4 model, also known as the Berkeley short-channel IGFET model (BSIM).[13–18] BSIM models are most commonly used in industry for modeling MOS digital circuits.

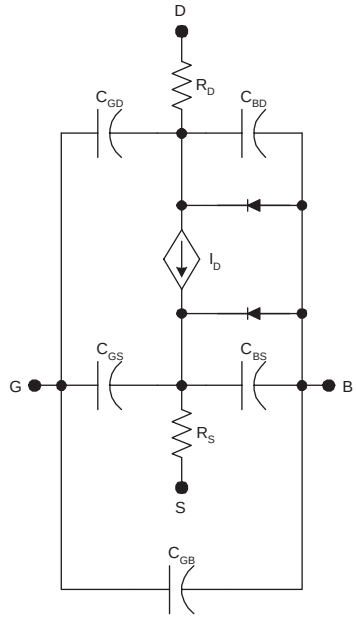


FIGURE 7.13
MOSFET SPICE model.

The examples in this book will use the level 1 model for clear illustration of the principles. However, an engineer who understands the use of the level 1 model can readily adapt to the use of the more complex models currently in use and under development.

All four types of SPICE MOSFET models are based on the circuit shown in Figure 7.13, where R_D and R_S are the parasitic resistances in the drain and source, respectively, C_{GD} is the capacitance between the gate and drain, C_{GS} is the capacitance between the gate and the source, etc. The diodes are parasitic p–n junctions between the body and the channel; they are always reverse biased under normal operating conditions.

The level 1 model uses the following equations to calculate the drain current in the MOSFET:

$$V_T = VTO + GAMMA\left(\sqrt{PHI + V_{SB}} - \sqrt{PHI}\right), \tag{7.45}$$

$$I_D = \frac{KP}{2}\frac{width}{length}\left(V_{GS} - V_T\right)^2\left(1 + V_{DS}LAMBDA\right) \text{ (saturation)}, \tag{7.46}$$

and

$$I_D = \frac{KP}{2}\frac{width}{length}\left[\left(V_{GS} - V_T\right)V_{DS} - \frac{V_{DS}^2}{2}\right]\left(1 + V_{DS}LAMBDA\right) \text{ (linear)}, \tag{7.47}$$

where

V_{GS} = gate-to-source voltage
V_{DS} = drain-to-source voltage
V_{SB} = source-to-body voltage
I_D = drain current source
V_{TO} = threshold voltage with zero body-to-source bias
$PHI = 2\phi_F$
width = gate width
length = gate length
KP = process transconductance parameter
LAMBDA = channel length modulation parameter

The device capacitances are calculated in the same fashion for all four SPICE models:

$$C_{GS} = width \times CGSO\,, \tag{7.48}$$

$$C_{GD} = width \times CGDO\,, \tag{7.49}$$

$$C_{GB} = width \times CGBO\,, \tag{7.50}$$

$$C_{BD} = width \times CBDO\,, \tag{7.51}$$

and

$$C_{BS} = width \times CBSO\,. \tag{7.52}$$

where

CGSO = gate-to source capacitance per unit gate width
CGDO = gate-to-drain capacitance per unit gate width
CGBO = gate-to-body capacitance per unit gate width
CBSO = body-to-source capacitance per unit gate width

SPICE MOSFET level 1 model parameters are listed in Table 7.2.

7.7 Integrated MOSFETs

Modern integrated MOSFETs are fabricated using self-alignment technology, which ensures minimum but guaranteed overlap of the gate with the source and drain. The gates are no longer made using metal, but with doped

TABLE 7.2

SPICE Level 1 Model Parameters for the MOSFET

Symbol	SPICE Name	Description	Units	Default	Typical
K'	KP	Process transconductance parameter	A/V^2	1E-4	1E-4
V_{TO}	VTO	Threshold voltage with $V_{BS} = 0$	V	0	0.5
γ	GAMMA	Body effect coefficient	$V^{1/2}$	0	1E-4
$2\phi_F$	PHI		V	1.0	1.0
C_{GSO}	CGSO	Gate–source capacitance per unit gate width	F/m	0	
C_{GDO}	CGDO	Gate–source capacitance per unit gate width	F/m	0	
C_{GBO}	CGBO	Gate–source capacitance per unit gate width	F/m	0	
C_{BSO}	CBSO	Gate–source capacitance per unit gate width	F/m	0	

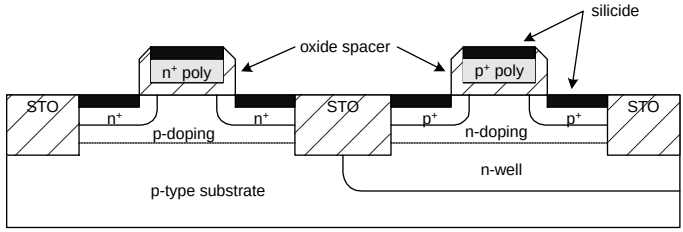


FIGURE 7.14
Integrated MOSFETs (STI = shallow trench isolation).

polysilicon instead.* However, metal is deposited on the polysilicon gates to reduce the parasitic gate resistance. Figure 7.14 shows integrated n-channel and p-channel MOSFETs fabricated on the same wafer using a typical CMOS process as described in Chapter 1.

7.8 PSPICE Simulations

Simulations were performed using PSPICE 9.1, which can be downloaded free.[19] The MOSFET level 1 model parameters used for these simulations are provided in Table 7.3. The common source characteristics (I_D vs. V_{DS} with

* In deep submicron MOSFETs, depletion of the polysilicon gate is becoming problematic. It is therefore likely that future device designs will call for metal gates once a suitable self-aligned technology is developed.

TABLE 7.3

MOSFET SPICE Parameters

Parameter	Value	Units
VTO	0.5	V
KP	50u	A/V^2
LAMBDA	0.05	—

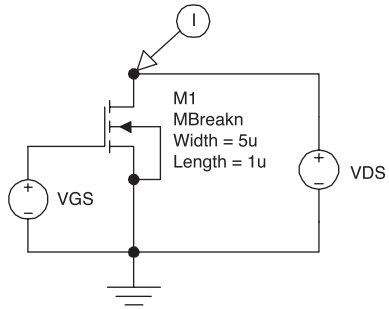


FIGURE 7.15
Circuit for the simulation of the characteristic curves of an n-channel MOSFET.

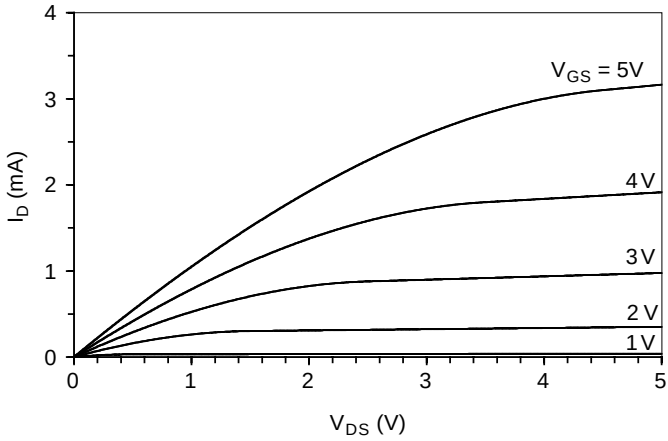


FIGURE 7.16
Simulated I_D vs. V_{DS} with V_{GS} as a parameter.

V_{GS} as a parameter) for a digital MOSFET were simulated using the circuit of Figure 7.15. The MOSFET common source characteristics appear in Figure 7.16. In the saturation region, the curves slope upward only slightly as a consequence of channel length modulation ($\lambda = 0.05$).

7.9 Summary

Field-effect transistors (FETs) are majority carrier devices and this makes them inherently fast. Three basic types of FETs are the metal oxide–semiconductor field-effect transistor (MOSFET), junction field-effect transistor (JFET), and metal–semiconductor field-effect transistor (MESFET). Of these, the MOSFET is most important in digital integrated circuits. FETs may be n-channel or p-channel; in n-channel FETS, conduction is entirely by electrons, while holes carry the current in p-channel FETs. FETS may be enhancement type (normally off) or depletion type (normally on). Normally off devices are preferred for digital integrated circuits.

For MOSFETs, important DC device parameters are the threshold voltage and the device transconductance parameter. The threshold voltage may be adjusted precisely by ion implantation. The device transconductance parameter is the product of the process transconductance parameter and the aspect ratio.

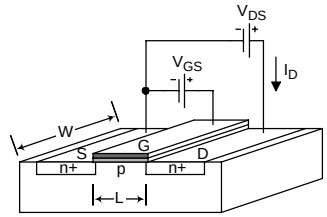
The three modes of operation for a MOSFET are cutoff, linear, and saturation. For cutoff operation, the device current is approximately zero and in saturation, the device acts like a voltage-controlled current source. (The gate–source voltage controls the drain current.) In linear operation, the device acts like a voltage-controlled resistance. (The gate–source voltage controls the drain-source resistance.) In some cases, the small drain current that flows under cutoff conditions must be considered. This subthreshold current increases exponentially with the gate-to-source bias of the MOSFET. Subthreshold conduction is important in determining the static dissipation in CMOS logic circuits.

Short-channel MOSFETs behave differently than long-channel transistors in several regards. Because the field lines under the gate terminate at the source and drain as well as the gate, the threshold voltage is a function of the gate length. This is called the short channel effect. Also, carrier velocity saturation tends to limit the drain currents to values less than those predicted based on the long-channel models. Additionally, channel length modulation is important in short-channel transistors so that the saturated drain current increases linearly with the drain-to-source bias.

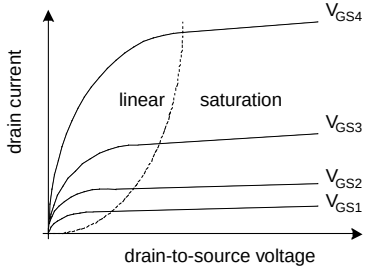
N-channel and p-channel MOSFETs are qualitatively similar in most regards. For the p-channel devices, however, all voltages and currents are opposite in polarity compared to the n-channel case.

FIELD-EFFECT TRANSISTORS QUICK REFERENCE

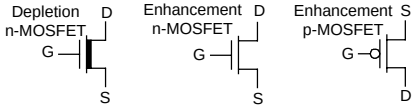
Enhancement n-MOSFET



n-MOSFET characteristics



MOSFET Symbols



FETs may be classified by gate structure (MOSFETs, MESFETs, JFETs), as enhancement-type or depletion type, and by channel carrier (n-channel or p-channel). FETs are majority carrier devices. FETs have three modes of operation: cutoff, linear (or ohmic), and saturation.

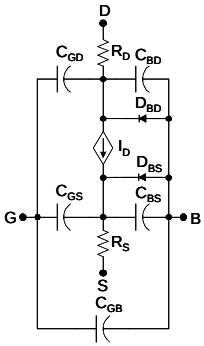
n-MOSFET Modes of Operation

cutoff	$V_{GS} < V_T$		$I_D \approx 0$
linear	$V_{GS} > V_T$	$V_{GS} - V_T > V_{DS}$	$I_D = K\left[(V_{GS} - V_T)V_{DS} - \frac{V_{DS}^2}{2}\right]$ $V_{DS} = (V_{GS} - V_T) - \sqrt{(V_{GS} - V_T)^2 - \frac{2I_D}{K}}$
saturation	$V_{GS} > V_T$	$V_{GS} - V_T < V_{DS}$	$I_D = \frac{K}{2}(V_{GS} - V_T)^2$

Device Transconductance Parameters

$$K_N = \frac{W_N}{L_N}k_N' \qquad k_N' = \frac{\mu_n \varepsilon_{OX}}{t_{OX}} \qquad K_P = \frac{W_P}{L_P}k_P' \qquad k_P' = \frac{\mu_p \varepsilon_{OX}}{t_{OX}}$$

SPICE Model



$$V_T = VTO + GAMMA\left(\sqrt{PHI + V_{SB}} - \sqrt{PHI}\right)$$

$$I_D = \frac{KP}{2}\frac{width}{length}(V_{GS} - V_T)^2(1 + V_{DS}LAMBDA)$$

$$I_D = \frac{KP}{2}\frac{width}{length}\left[(V_{GS} - V_T)V_{DS} - \frac{V_{DS}^2}{2}\right](1 + V_{DS}LAMBDA)$$

$$C_{GS} = width \times CGSO \qquad C_{GD} = width \times CGDO \qquad C_{GB} = width \times CGBO$$

$$C_{BD} = width \times CBDO \qquad C_{BS} = width \times CBSO$$

Design Rules

$\mu_n = 580 cm^2 V^{-1} s^{-1}$ $\mu_p = 230 cm^2 V^{-1} s^{-1}$ $\varepsilon_{OX} = 3.9\varepsilon_0$ Å = 10^{-8} cm

f = 10^{-15} p = 10^{-12} n = 10^{-9} μ = 10^{-6} m = 10^{-3} k = 10^{3} M = 10^{6} G = 10^{9}

Laboratory Exercises

L7.1. For a commercially available n-MOSFET, obtain the datasheet from the manufacturer's Web site. Measure I_D vs. V_{GS} with V_{DS} constant at one half the absolute maximum rating. From the measured characteristic, estimate the threshold voltage and the device transconductance parameter. Compare your values to the manufacturer's specifications on the datasheet.

L7.2. Repeat L7.1 for a commercially available p-MOSFET.

L7.3. For a commercially available n-MOSFET, measure and plot the characteristic curves. Vary the drain current up to one half of the rated absolute maximum from the datasheet. Vary the drain-source voltage up to one half of the rated absolute maximum from the datasheet. Using the characteristic curves, plot the channel length modulation parameter vs. the drain current. Are your results consistent with the specifications given in the datasheet?

Problems

P7.1. Calculate the zero-bias threshold voltage for an n-channel MOSFET with t_{OX} = 6 nm and $N_a = 1.2 \times 10^{16}$ cm^{-3}. Assume that the gate is heavily-doped n-polysilicon (with the Fermi level coincident with the conduction band) and that 10^{11} cm^{-2} positive charges are in the oxide. A boron dose of 10^{12} cm^{-2} is implanted to adjust the threshold voltage.

P7.2. Calculate the zero-bias threshold voltage for a p-channel MOSFET with t_{OX} = 6 nm and $N_d = 10^{16}$ cm^{-3}. Assume that the gate is heavily doped p-polysilicon (with the Fermi level coincident with the valence band) and that 10^{11} cm^{-2} positive charges are in the oxide. A phosphorus dose of 5×10^{11} cm^{-2} is implanted to adjust the threshold voltage.

P7.3. Consider an n-channel MOSFET with t_{OX} = 7 nm and $N_a = 10^{16}$ cm^{-3}. Determine the required ion implantation (impurity and dose) to adjust the threshold by +0.45 V.

P7.4. Consider an n-channel MOSFET with t_{OX} = 7 nm and $N_a = 10^{16}$ cm^{-3}. Assume that the gate is heavily doped n-polysilicon (with the Fermi level coincident with the conduction band) and that 10^{11} cm^{-2} positive charges are in the oxide. Determine the required ion implantation (impurity and dose) to adjust the threshold voltage to 0.4 V.

P7.5. Consider n-channel and p-channel silicon MOSFETs fabricated on the same wafer with channel lengths of 0.5 μm and 20-nm thick silicon dioxide.

1. Determine the process transconductance parameters for n-channel and p-channel devices.
2. Determine the required aspect ratios for n-channel and p-channel MOSFETs such that the device transconductance parameters are both 0.5 mA/V^2.
3. Determine the oxide capacitances of the devices. (The oxide capacitance is approximately the parallel plate capacitance for plates of area W × L and separation t_{OX}.)

P7.6. Consider n-channel and p-channel silicon MOSFETs fabricated on the same wafer with channel lengths of 0.3 μm and 14-nm thick silicon dioxide.

1. Determine the process transconductance parameters for n-channel and p-channel devices.
2. Determine the required aspect ratios for n-channel and p-channel MOSFETs such that the device transconductance parameters are both 0.5 mA/V^2.
3. Determine the oxide capacitances of the devices. (The oxide capacitance is approximately the parallel plate capacitance for plates of area W × L and separation t_{OX}.)

P7.7. Calculate and plot the characteristic curves (I_D vs. V_{DS} with V_{GS} as a parameter) for an n-channel MOSFET with $W = 2$ μm, $L = 0.25$ μm, $t_{OX} = 6$ nm, and $V_T = 0.4$ V. Consider $0 \leq V_{DS} \leq 2.5$ V and $V_{GS} = 0$, 0.5, 1, 1.5, 2, and 2.5 V.

P7.8. Calculate and plot the subthreshold current for the following n-channel MOSFETs:

1. A bulk n-channel MOSFET with $K = 1.6$ mA/V^2, $V_T = 0.3$ V, and $m = 1.6$
2. A silicon-on-insulator (SOI) n-channel MOSFET with $K = 1.6$ mA/V^2, $V_T = 0.3$ V, and $m = 1$.

P7.9. The *destructive* breakdown of silicon dioxide occurs with an electric field of about 10 million V/cm. Consider an integrated MOSFET with a 4.5-nm thick oxide.

1. Determine the gate-to-source bias voltage that will result in destructive breakdown.
2. Determine the associated charge in Coulombs for a 0.18 μm × 2 μm gate.
3. Determine the associated number of electrons.

P7.10. Estimate the transit time for an n-channel MOSFET with L = 100 nm and $V_{DS} = 1.5$ V, assuming

1. The constant mobility model
2. The velocity saturation model

Which model is more appropriate in this case?

References

1. Sze, S.M., *Physics of Semiconductor Devices*, 2nd ed., John Wiley & Sons, New York, 1981.
2. Streetman, B.G. and Banerjee, S., *Solid State Electronic Devices*, Prentice Hall, Englewood Cliffs, NJ, 1999.
3. Long, S.I. and Butner, S.E., *Gallium Arsenide Digital Integrated Circuit Design*, McGraw-Hill, New York, 1990.
4. Nicollian, E.H. and Brews, J.R., *MOS Physics and Technology*, John Wiley & Sons, New York, 1982.
5. Brews, J.R., *Physics of the MOS Transistor, Applied Solid State Science*, Supplement 2A, Kahng, D., Ed., Academic Press, New York, 1981.
6. Taur, Y. and Ning, T.H., *Fundamentals of Modern VLSI Devices*, Cambridge University Press, New York, 1998.
7. Brews, J.R., A charge sheet model of the MOSFET, *Solid-State Electron.*, 21, 345, 1978.
8. Nguyen, T.N. and Plummer, J.D., Physical mechanisms responsible for short-channel effects in MOS devices, *Tech. Dig.1981 Int. Electron. Devices Meet.*, 596, 1981.
9. Yau, L.D., A simple theory to predict the threshold voltage of short-channel IGFETs, *Solid-State Electron.*, 17, 1059, 1974.
10. Taur, Y., Hsu, C.H., Wu, B., Kiehl, R., Davari, B., and Shahidi, G., Saturation transconductance of deep submicron-channel MOSFETs, *Solid-State Electron.*, 36, 1085, 1993.
11. Caughey, D.M. and Thomas, R.E., Carrier mobilities in silicon empirically related to doping and field, *Proc. IEEE*, 55, 2192, 1967.
12. Taylor, G.W., Velocity saturated characteristics of short-channel MOSFETs, *Bell Labs. Tech. J.*, 63, 1325, 1984.
13. Pang, Y.-S. and Brews, J.R., Models for subthreshold and above-threshold currents in 0.1-μm pocket n-MOSFETs for low-voltage applications, *IEEE Trans. Electron. Devices*, 49, 832, 2002.
14. Hu, C., BSIM model for circuit design using advanced technologies, *Dig. Tech. Papers 2001 Symp. VLSI Circuits*, 5, 2001.
15. Cheng, Y., Jeng, M.-C., Liu, Z., Huang, J., Chan, M., Chen, K., Ko, P.K., and Hu, C., A physical and scalable I-V model in BSIM3v3 for analog/digital circuit simulation, *IEEE Trans. Electron. Devices*, 277, 1997.
16. Gowda, S.M. and Sheu, B.J., BSIM plus: an advanced SPICE model for submicron MOS VLSI circuits, *IEEE Trans. Computer-Aided Design Integrated Circ. Syst.*, 13, 1166, 1994.
17. Arora, N., *MOSFET Models for VLSI Circuit Simulation*, Springer-Verlag, Vienna, 1993.
18. Sheu, B.J., Scharfetter, D.L., Ko, P.K., and Jeng, M.C., BSIM Berkeley short-channel IGFET model, *IEEE J. Solid-State Circuits*, 22, 558, 1987.
19. www.cadence.com (Cadence).

8

NMOS Logic

8.1 Introduction

NMOS logic circuits use only n-channel MOSFETs as active devices. Because of the small size of MOSFETs, NMOS integrated circuits can be made with the highest possible packing density. The only disadvantage of NMOS is the high standby power, which can be comparable to that for bipolar circuits. NMOS was once used extensively for logic circuits, including early microprocessors; however, CMOS has replaced NMOS in most of these applications today. Nonetheless, NMOS remains important in state-of-the-art memories. High-density dynamic random access memory chips use one-transistor storage cells comprising a single n-channel MOSFET and a single capacitor per bit.

8.2 Circuit Evolution

NMOS circuits comprise a load device and one or more NMOS switches. The simplest form uses a resistor as the pull-up device. Other circuit designs use pull-up transistors for improved circuit density. An NMOS inverter with a resistor pull-up is shown in Figure 8.1. This circuit is essentially a MOSFET common source amplifier overdriven between cutoff and ohmic operation. The basic operation is as follows. With a logic-zero input, the MOSFET is cut off and the output rises to the supply voltage V_{DD}. With a logic-one input, the MOSFET is driven to ohmic (linear) operation. Under this condition, the output voltage is less than the threshold voltage for the switch MOSFETs in the next stage. The primary disadvantage of this circuit design is the low packing density. Typically, the resistors take up far more chip area than the switch transistors, making the design very inefficient.

The packing density of NMOS circuits can be greatly improved by replacing the pull-up resistor with a MOSFET. If the pull-up is an enhancement type device, it may be fabricated with the same threshold voltage as the

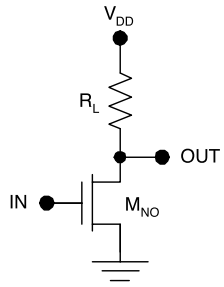


FIGURE 8.1
NMOS inverter with resistor pull-up.

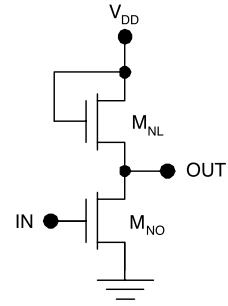


FIGURE 8.2
NMOS inverter with an enhancement-type pull-up device.

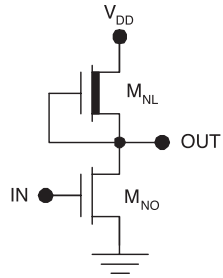


FIGURE 8.3
NMOS inverter with a depletion-type pull-up device.

switch transistors. Using a single type of transistor minimizes the number of fabrication process steps and therefore the cost. Figure 8.2 shows such a circuit with an enhancement-type n-MOSFET pull-up transistor. A disadvantage of this design relative to the resistor pull-up is reduced voltage swing because the maximum output voltage is $V_{DD} - V_T$ rather than V_{DD}, as in the case of the resistor pull-up.

An improved NMOS circuit uses a depletion type pull-up transistor to achieve high packing density and maximum voltage swing. This is the most commonly used approach (Figure 8.3) and will be discussed in detail in the following sections.

8.3 Voltage Transfer Characteristic

Consider the depletion load NMOS inverter shown in Figure 8.4. With a logic-zero input, the switch transistor, M_{NO}, is cut off. For the load device, M_{NL}, the drain current is 0 and the device is in the ohmic region of operation. Therefore, the drain-to-source voltage for the load is 0 and

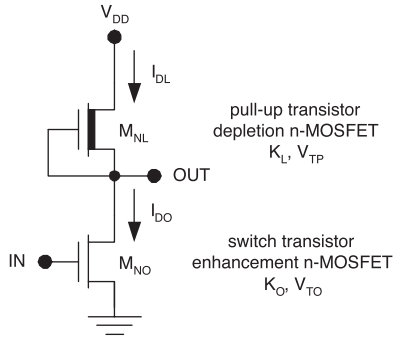


FIGURE 8.4
NMOS inverter.

$$V_{OH} = V_{DD}. \tag{8.1}$$

With a logic-one input, the switch transistor operates in the ohmic region while the load transistor is saturated. Equating the drain currents for the two devices yields

$$K_O\left[\left(V_{DD} - V_{TO}\right)V_{OL} - \frac{V_{OL}^2}{2}\right] = \frac{K_L}{2}V_{TL}^2, \tag{8.2}$$

where K_O and K_L are the device transconductance parameters for M_{NO} and M_{NL}, respectively, V_{TO} and V_{TL} are the threshold voltage for M_{NO} and M_{NL}, respectively, and it has been assumed that $V_{IN} = V_{DD}$ (V_{OH} from a similar gate). Solving for V_{OL} yields

$$V_{OL} = V_{DD} - V_{TO} \pm \sqrt{\left(V_{DD} - V_{TO}\right)^2 - \left(\frac{K_L}{K_O}\right)V_{TL}^2}. \tag{8.3}$$

Notice that the quadratic formula predicts two solutions; however, the solution stemming from use of the plus sign is nonphysical and must be discarded. It can be concluded that the output low voltage depends on the ratio of the device transconductance parameters but not their absolute values. Thus, both devices can be scaled up or down in size without affecting the output low voltage.

The input low voltage for the depletion load NMOS gate can be determined with the assumption that M_{NO} is saturated and M_{NL} is linear. Then the drain currents can be written as

$$I_{DO} = \frac{K_O}{2}\left(V_{\text{IN}} - V_{TO}\right)^2 \tag{8.4}$$

and

$$I_{DL} = K_L\left[\left(-V_{TL}\right)\left(V_{DD} - V_{\text{OUT}}\right) - \frac{\left(V_{DD} - V_{\text{OUT}}\right)^2}{2}\right]. \tag{8.5}$$

If the drain currents are equated, then

$$I_{DO} = I_{DL}, \tag{8.6}$$

$$dI_{DO} = dI_{DL}, \tag{8.7}$$

and

$$\frac{\partial I_{DO}}{\partial V_{\text{IN}}} dV_{\text{IN}} = \frac{\partial I_{DL}}{\partial V_{\text{OUT}}} dV_{\text{OUT}}. \tag{8.8}$$

The slope of the transfer characteristic can therefore be determined using the partial derivatives:

$$\frac{dV_{\text{OUT}}}{dV_{\text{IN}}} = \frac{\dfrac{\partial I_{DO}}{\partial V_{\text{IN}}}}{\dfrac{\partial I_{DL}}{\partial V_{\text{OUT}}}} = \frac{K_O\left(V_{\text{IN}} - V_{TO}\right)}{K_L V_{TL} + K_L\left(V_{DD} - V_{\text{OUT}}\right)}. \tag{8.9}$$

By definition, this slope is –1 at the input low voltage. Therefore,

$$\left.\frac{K_O\left(V_{IN} - V_{TO}\right)}{K_L V_{TL} + K_L\left(V_{DD} - V_{\text{OUT}}\right)}\right|_{V_{\text{IN}} = V_{IL}} = -1. \tag{8.10}$$

Solving for V_{OUT} yields

$$V_{\text{OUT}} = \frac{K_O}{K_L}\left(V_{\text{IN}} - V_{TO}\right) + V_{TL} + V_{DD}; \tag{8.11}$$

substituting this result into Equation 8.6 and solving it yields the input low voltage:

$$V_{IL} = V_{TO} + \frac{K_L}{\sqrt{K_O K_L + K_O^2}}\left|V_{TL}\right|. \tag{8.12}$$

A similar approach can be used for the determination of V_{IH}. In this case, start with the assumption that M_{NO} is linear and M_{NL} is saturated. Then the drain currents are given by

$$I_{DO} = K_O\left[\left(V_{\text{IN}} - V_{TO}\right)V_{\text{OUT}} - \frac{V_{\text{OUT}}^2}{2}\right] \tag{8.13}$$

and

$$I_{DL} = \frac{K_L V_T^2}{2}\,. \tag{8.14}$$

The drain currents are equal

$$I_{DO} = I_{DL}\,, \tag{8.15}$$

so

$$dI_{DO} = dI_{DL}\,. \tag{8.16}$$

However, I_{DL} is constant, so

$$\frac{\partial I_{DL}}{\partial V_{\text{IN}}} = \frac{\partial I_{DL}}{\partial V_{\text{OUT}}} = 0\,. \tag{8.17}$$

Therefore,

$$\frac{\partial I_{DO}}{\partial V_{\text{IN}}}dV_{\text{IN}} + \frac{\partial I_{DO}}{\partial V_{\text{OUT}}}dV_{\text{OUT}} = 0 \tag{8.18}$$

and the slope of the voltage transfer characteristic can be found from

$$\frac{dV_{\text{OUT}}}{dV_{\text{IN}}} = \frac{\dfrac{\partial I_{DO}}{\partial V_{\text{IN}}}}{-\dfrac{\partial I_{DO}}{\partial V_{\text{OUT}}}}\,. \tag{8.19}$$

By definition, the input high voltage is the value of input voltage for which the slope of the voltage transfer characteristic is –1. Using this condition and solving the equation yields

$$V_{IH} = V_{TO} + 2\left|V_{TL}\right|\sqrt{\frac{K_L}{3K_O}}\,. \tag{8.20}$$

TABLE 8.1

Parameters of Example Depletion Load NMOS Inverter

Device	L (μm)	W (μm)	K (mA/V²)	V_T (V)
M_{NL}	0.5	12.0	0.69	0.6
M_{NO}	0.5	1.0	0.057	–0.3

Example 8.1

Design a depletion load NMOS inverter with V_{DD} = 2.5 V, V_{TO} = 0.6 V, and V_{TL} = –0.3 V. The fabrication process uses 0.5-μm technology with t_{ox} = 7 nm and μ_n = 580 cm²/V.

Solution. The transistors must be designed so that V_{OL} is several tenths of a volt less than V_{TO}. Designing for V_{OL} = 0.3 V, the required ratio of device transconductance parameters is

$$\frac{K_O}{K_L} = \frac{V_{TL}^2}{2\left[\left(V_{DD} - V_{TO}\right)V_{OL} - \frac{V_{OL}^2}{2}\right]} = \frac{(-0.3\text{ V})^2}{2\left[(2.5\text{ V} - 0.6\text{ V})0.3\text{ V} - \frac{(0.3\text{ V})^2}{2}\right]} \approx \frac{1}{12}.$$

Therefore,

$$\frac{W_O/L_O}{W_L/L_L} \approx \frac{1}{12},$$

where W_O and L_O are the width and length of M_{NO}, respectively, and W_L and L_L are the width and length of the load device, M_{NL}, respectively. However, both devices should use the minimum gate length for the best dynamic performance.

Example 8.2

Determine the critical voltages of the transfer function for the NMOS inverter shown in Figure 8.5.

Solution. The device transconductance parameters are given by

$$K_L = \left(\frac{5\ \mu\text{m}}{0.5\ \mu\text{m}}\right)220\ \mu\text{A/V}^2 = 2.2\ \text{mA/V}^2$$

and

$$K_O = \left(\frac{8\ \mu\text{m}}{0.5\ \mu\text{m}}\right)220\ \mu\text{A/V}^2 = 3.5\ \text{mA/V}^2.$$

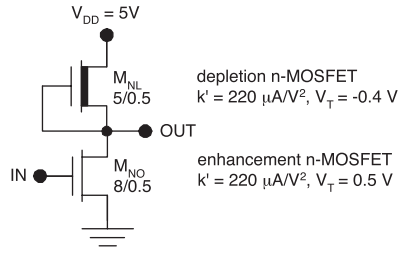


FIGURE 8.5
Example NMOS inverter for the calculation of the VTC.

The four critical voltages are

$$V_{IL} = V_{TO} + \frac{K_L}{\sqrt{K_O K_L + K_O^2}}\left|V_{TL}\right|$$

$$= 0.5\ \text{V} + \frac{2.2\ \text{mA/V}^2}{\sqrt{\left(3.5\ \text{mA/V}^2\right)\left(2.2\ \text{mA/V}^2\right) + \left(3.5\ \text{mA/V}^2\right)^2}}\left|-0.4\ \text{V}\right| = 0.70\ \text{V},$$

$$V_{IH} = V_{TO} + 2\left|V_{TL}\right|\sqrt{\frac{K_L}{3K_O}} = 0.5\ \text{V} + 2\left|-0.4\ \text{V}\right|\sqrt{\frac{2.2\ \text{mA/V}^2}{3\left(3.5\ \text{mA/V}^2\right)}} = 0.87\ \text{V},$$

$$V_{OL} = V_{DD} - V_{TO} \pm \sqrt{\left(V_{DD} - V_{TO}\right)^2 - \left(\frac{K_L}{K_O}\right)V_{TL}^2}$$

$$= 5\ \text{V} - 0.5\ \text{V} \pm \sqrt{(5\ \text{V} - 0.5\ \text{V})^2 - \left(\frac{2.2\ \text{mA/V}^2}{3.5\ \text{mA/V}^2}\right)(-0.4\ \text{V})^2} = 0.0112\ \text{V},$$

and

$$V_{OH} = V_{DD} = 5\ \text{V}.$$

Example 8.3

Determine the voltage transfer characteristic for the NMOS gate of Figure 8.6.

Solution. The voltage transfer characteristic may be determined point by point by equating the drain currents of the transistors and solving for the output voltage. This requires knowledge of the operating modes for the transistors at each point, as can be determined with the aid of the rules summarized in Table 8.2. For $V_{IN} \le 0.5$ V, M_{NO} is cut off and M_{NL} is linear, so

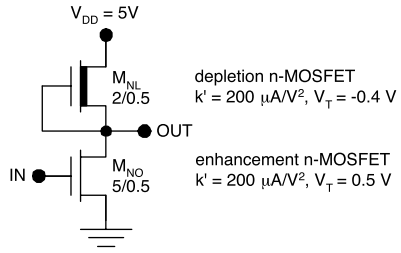


FIGURE 8.6
Example NMOS inverter.

TABLE 8.2

Modes of Operation for n-Channel MOSFETs

Cutoff	$V_{GS} \le V_T$
Saturation	$V_T \le V_{GS} \le (V_{DS} + V_T)$
Linear	$[V_{GS} \ge V_T]$ and $[V_{GS} \ge (V_{DS} + V_T)]$

$$V_{\text{OUT}} = 5 \text{ V} \quad (V_{\text{IN}} \le 0.5 \text{ V}) .$$

If $(V_{\text{OUT}} + 0.5 \text{ V}) \le V_{\text{IN}} \le 0.5$ V and $V_{\text{OUT}} \ge 4.7$ V, M_{NO} is saturated and M_{NL} is linear, so

$$I_{DD} = \frac{K_O}{2}\left(V_{\text{IN}} - V_{TO}\right)^2$$

and

$$V_{\text{OUT}} = V_{DD} - V_{DSL} = V_{DD} - \left[-V_{TL} - \sqrt{\left(-V_{TL}\right)^2 - \frac{K_O\left(V_{\text{IN}} - V_{TO}\right)^2}{K_L}} \right]$$

$$= 4.7 \text{ V} + \sqrt{(0.3 \text{ V})^2 - (1.25)\left(V_{\text{IN}} - 0.5 \text{ V}\right)^2} ;$$

$$\left[\left(V_{\text{OUT}} + 0.5 \text{ V}\right) \le V_{\text{IN}} \le 0.5 \text{ V}\right] \text{ and } \left[V_{\text{OUT}} \ge 4.7 \text{ V}\right]$$

Then, if M_{NO} is linear and M_{NL} is saturated so that

$$I_{DD} = \frac{K_L}{2}\left(-V_{TL}\right)^2$$

and

$$V_{\text{OUT}} = V_{DSO} = \left(V_{\text{IN}} - V_{TO}\right) - \sqrt{\left(V_{\text{IN}} - V_{TO}\right)^2 - \frac{K_L\left(-V_{TL}\right)^2}{K_O}} -$$

$$= V_{IN} - 0.5\ \text{V} - \sqrt{\left(V_{IN} - 0.5\ \text{V}\right)^2 - (-0.3)^2 / 1.25};$$

$$\left[\left(V_{OUT} + 0.5\ \text{V}\right) \geq V_{IN}\right] \text{and} \left[V_{OUT} \leq 4.7\ \text{V}\right].$$

Figure 8.7 shows the voltage transfer characteristic and Table 8.3 provides some representative points.

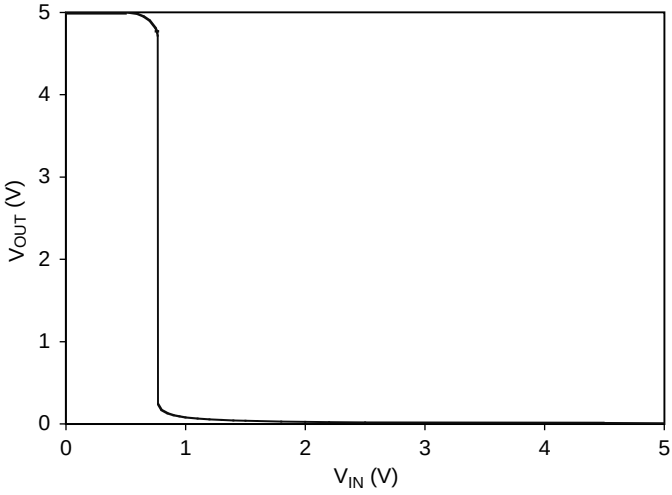


FIGURE 8.7
Voltage transfer characteristic of example NMOS gate.

TABLE 8.3

Some Representative Points from the Voltage Transfer Characteristic of the Example NMOS Gate

V_{IN} (V)	V_{OUT} (V)	M_{NO}	V_{GSO}–V_{TO}–V_{DSO} (V)	M_{NL}	V_{GSL}–V_{TL}–V_{DSL} (V)
0.00	5.00	CO[a]	–5.50	LIN	0.300
0.50	5.00	CO	–5.00	LIN	0.300
0.70	4.90	SAT[b]	–4.70	LIN	0.200
0.75	4.81	SAT	–4.56	LIN	0.109
0.768	4.70	SAT	–4.44	LIN	0.0148
0.80	0.166	LIN[c]	0.1342	SAT	–4.53
0.90	0.103	LIN	0.297	SAT	–4.60
1.00	0.078	LIN	0.422	SAT	–4.62
1.5	0.037	LIN	0.963	SAT	–4.66
2.00	0.024	LIN	1.476	SAT	–4.68
3.00	0.014	LIN	2.49	SAT	–4.69
4.00	0.010	LIN	3.49	SAT	–4.69
5.00	0.008	LIN	4.49	SAT	–4.69

[a] CO = cutoff.
[b] SAT = saturated.
[c] LIN = linear.

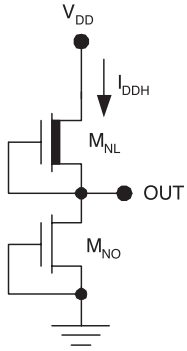


FIGURE 8.8
NMOS inverter for the determination of P_H.

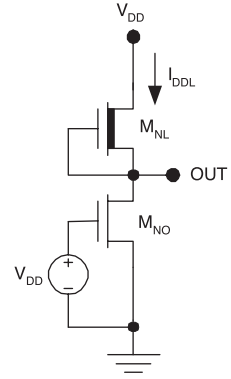


FIGURE 8.9
NMOS inverter for the determination of P_L.

8.4 Dissipation

The two contributions to the power dissipation in an NMOS gate are the DC (static) and AC (dynamic) components. Usually, the static dissipation dominates in NMOS logic, but this depends on the switching frequency. The power dissipation should be minimized to extend battery life and simplify chip cooling requirements; however, there is always a trade-off between circuit speed and power.

The DC or static dissipation depends on the output state of the gate. Consider the depletion load NMOS inverter with the output high as illustrated in Figure 8.8 (the input is grounded). With a logic-one output, the switch transistor is cut off so that the supply current, I_{DDH}, is approximately 0 (apart from the small leakage current). Therefore,

$$P_H = V_{DD} I_{DDH} \approx 0 \,. \tag{8.21}$$

With a logic-zero output as shown in Figure 8.9, the load transistor, M_{NL}, is saturated and the switch transistor, M_{NO}, is linear. Therefore, the output low supply current is

$$I_{DDL} = \frac{K_L V_{TL}^2}{2} \tag{8.22}$$

and the static dissipation with the output low is

$$P_L = V_{DD} I_{DDL} = \frac{K_L V_{DD} V_{TL}^2}{2} \,. \tag{8.23}$$

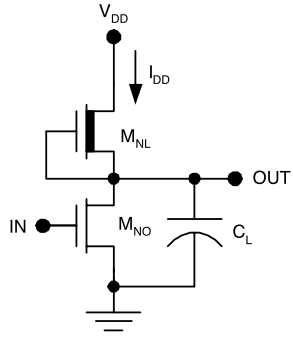


FIGURE 8.10
NMOS inverter with a lumped capacitive load for determination of the dynamic dissipation.

The average static dissipation depends on the output duty cycle, but it is common practice to assume a 50% duty cycle. Therefore, the average static dissipation is

$$P_{DC} \approx \frac{P_H + P_L}{2} = \frac{K_L V_{DD} V_{TL}^2}{4} . \tag{8.24}$$

The DC dissipation can be reduced by a reduction of the supply voltage or by scaling down the K values. (Although K_O does not appear explicitly in the power equation, the two K values are scaled up or down together in order to preserve the desired voltage transfer characteristic.) Scaling down the K values degrades the switching speed of the circuitry unless the load capacitances can be scaled similarly. The dynamic or AC dissipation is associated with the switching of the output. Consider the NMOS inverter with a lumped capacitive load as shown in Figure 8.10.

The energy associated with one switching cycle (a low-to-high transition at the output, followed by a high-to-low transition at the output) is

$$J = V_{DD} \int_{\substack{clock \\ cycle}} i_{DD} dt . \tag{8.25}$$

If the current that flows in M_{NO} during the low-to-high transition is neglected, then

$$i_{DD} = C_L \frac{dV_{\text{OUT}}}{dt} , \tag{8.26}$$

so

$$J = V_{DD} \int_0^{V_{DD}} C_L dV_{\text{OUT}} = C_L V_{DD}^2 , \tag{8.27}$$

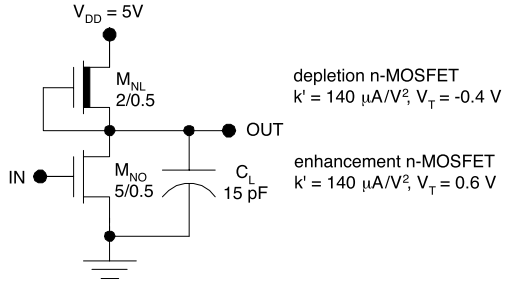


FIGURE 8.11
NMOS inverter for the example calculation of dissipation.

which is the energy, in joules, dissipated during each switching cycle. The dynamic dissipation is therefore

$$P_{AC} = fC_L V_{DD}^2 = \alpha f_{CLK} C_L V_{DD}^2, \tag{8.28}$$

where f is the switching frequency, f_{CLK} is the clock frequency, and α is the switching activity. The switching activity is less than unity, so the actual switching frequency for any particular gate will be less than the system clock frequency. The overall dissipation is the sum of the static and dynamic components,

$$P = P_{DC} + P_{AC} = \frac{K_L V_{DD} V_{TL}^2}{4} + fC_L V_{DD}^2. \tag{8.29}$$

Example 8.4

Calculate the dissipation for the NMOS inverter with a 15-pF load as shown in Figure 8.11.

Solution. The static dissipation is

$$P_{DC} \approx \frac{P_H + P_L}{2} = \frac{K_L V_{DD} V_{TL}^2}{4} = \frac{(0.56 \text{ mA/V}^2)(5V)(-0.4 \text{ V})^2}{4} = 0.112 \text{ mW};$$

the dynamic dissipation is given by $P_{AC} = fC_L V_{DD}^2 = f(15 \text{ pf})(5 \text{ V})^2 = f(0.375 \text{ nJ})$ and the total dissipation is $P = P_{DC} + P_{AC} = 0.112 \text{ mW} + f(0.375 \text{ nJ})$. The dynamic and static dissipation components will be equal at a switching frequency of 300 kHz.

8.5 Propagation Delays

To estimate the propagation delays of a depletion load NMOS inverter, consider a lumped capacitive load as shown in Figure 8.12. Real loads

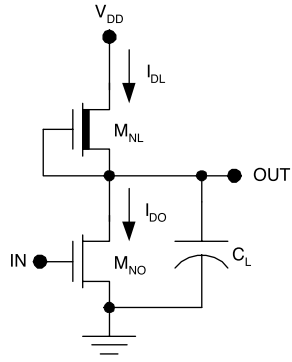


FIGURE 8.12
NMOS inverter with a lumped capacitive load.

involve distributed capacitances, resistances, and inductances, so the lumped capacitive load is an approximation.

The low-to-high propagation delay, t_{PLH}, can be estimated as follows. Assume that the fall time at the input is negligible. (Although the output makes a low-to-high transition, the input makes a high-to-low transition.) At $t = 0$, the input voltage decreases abruptly to cut off the switch transistor, M_{NO}. M_{NL} is saturated and the drain current is given by

$$I_{DL} = \frac{K_L V_{TL}^2}{2}; \tag{8.30}$$

therefore, the time derivative of the output voltage is

$$\frac{dV_{\text{OUT}}}{dt} = \frac{1}{C_L}\frac{K_L V_{TL}^2}{2}. \tag{8.31}$$

By definition of the propagation delay, V_{OUT} reaches $V_{DD}/2$ at $t = t_{PLH}$, thus solving

$$t_{PLH} = \frac{V_{DD} C_L}{K_L V_{TL}^2}. \tag{8.32}$$

This shows that the low-to-high propagation delay is proportional to the load capacitance and inversely proportional to the current driving capability of the load transistor.

The high-to-low propagation delay can be estimated using the assumption that the rise time at the input is negligible. (Although the output makes a high-to-low transition, the input makes a low-to-high transition.) At $t = 0^+$, M_{NO} will be saturated and M_{NL} will be linear. Thus at $t = 0^+$, the drain currents are given by

$$I_{DO} = \frac{K_O}{2}\left(V_{DD} - V_{TO}\right)^2 \tag{8.33}$$

and

$$I_{DL} = K_L\left[\left|V_{TL}\right|\left(V_{DD} - V_{\text{OUT}}\right) - \frac{\left(V_{DD} - V_{\text{OUT}}\right)^2}{2}\right]. \tag{8.34}$$

The drain current in the load transistor depends on the square of the output voltage, resulting in a nonlinear differential equation. However, the analysis can be greatly simplified by neglecting I_{DL}; very little accuracy is lost. This approximation yields

$$t_{PHL} \approx \frac{V_{DD}C_L}{K_O\left(V_{DD} - V_{TO}\right)^2}. \tag{8.35}$$

This shows that the high-to-low propagation delay is directly proportional to the load capacitance and inversely proportional to the current driving capability of the switch transistor. As a first approximation, the propagation delay is also inversely proportional to the supply voltage because the squared term in the denominator is dominant.

Example 8.5

Estimate the propagation delays for the NMOS inverter shown in Figure 8.13.

Solution. The low-to-high propagation delay is

$$t_{PLH} = \frac{V_{DD}C_L}{K_L V_{TL}^2} = \frac{(5\text{ V})(15\text{ pF})}{\left(0.56\text{ mA/V}^2\right)(-0.4\text{ V})^2} = 840\text{ ns}$$

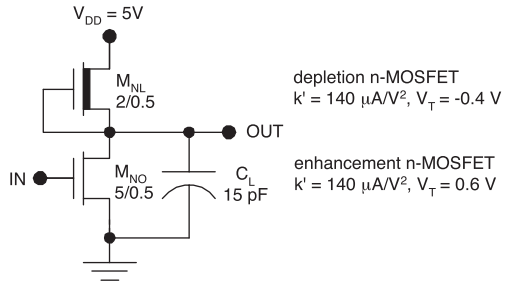


FIGURE 8.13
Example of NMOS inverter for calculation of the propagation delays.

and the high-to-low propagation delay is

$$t_{PHL} \approx \frac{V_{DD}C_L}{K_O\left(V_{DD}-V_{TO}\right)^2} = \frac{(5\text{ V})(15\text{ pF})}{\left(1.4\text{ mA/V}^2\right)(5\text{ V}-0.6\text{ V})^2} = 2.8\text{ ns}\,.$$

Therefore, t_{PLH} is 300 times longer than t_{PHL}. In order to make the propagation delays more nearly equal, it would be necessary to increase K_L/K_O or to make V_{TL} more negative. Either modification would increase V_{OL}. These choices need to be made with due consideration of the DC voltage transfer characteristic.

8.6 Fan-Out

The fan-out of NMOS gates is determined by dynamic rather than DC (static) considerations. Suppose that an NMOS gate is loaded with similar NMOS circuits. The only DC load current is due to the leakage currents in the MOS gates of the switch transistors and this is negligible. However, the load gates present a significant capacitive load. Therefore, the propagation delays increase with the fan-out and the maximum fan-out is dictated by the maximum tolerable propagation delay.*

Consider a depletion load NMOS gate loaded by N similar NMOS gate circuits as shown in Figure 8.14. The maximum fan-out can be estimated based on the following assumptions: 1) the maximum allowable propagation delay, $t_{P,\max}$, has been determined based on system engineering considerations; 2) the low-to-high propagation delay is much longer than the high-to-low propagation delay; and 3) the fan-out gates can be treated as a lumped capacitive load equal to NC_{OX}. Then,

$$C_L \le \frac{K_L V_{TL}^2 t_{P,\max}}{V_{DD}} \tag{8.36}$$

and

$$N \le \frac{K_L V_{TL}^2 t_{P,\max}}{V_{DD}C_{IN}}\,, \tag{8.37}$$

where C_{IN} is the input capacitance for the NMOS load gates. The input capacitance can be estimated by the parallel plate capacitance of the MOS structure, so

* As a rule of thumb, the maximum propagation delay should be less than 5% of the clock period.

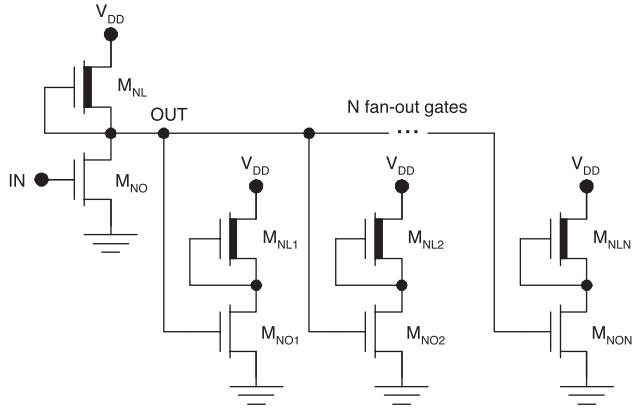


FIGURE 8.14
NMOS inverter with N similar fan-out gates.

$$C_{IN} \approx \frac{\varepsilon_{ox} W_O L_O}{t_{ox}}, \tag{8.38}$$

where ε_{ox} is the permittivity of silicon dioxide, 3.45×10^{-13} F/cm, W_O and L_O are the width and length of the gates in the switch transistors, and t_{ox} is the gate oxide thickness.

The maximum fan-out, N_{MAX}, is the largest integer satisfying Equation 8.37. For low-frequency applications the maximum fan-out of an NMOS gate may be in the hundreds or even thousands. For higher-frequency applications, the fan-out should be at least 10. Smaller values greatly restrict the logic design of the system.

Example 8.6

Estimate the maximum fan-out for the NMOS design illustrated in Figure 8.15 if the system clock frequency is to be 100 MHz. V_{TO} = 0.6 V and V_{TL} = –0.5 V.

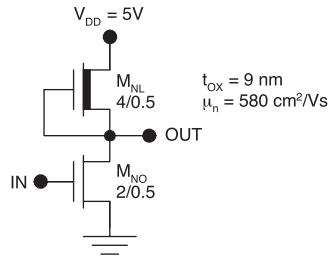


FIGURE 8.15
Example of NMOS inverter for calculation of the maximum fan-out.

Solution. The device transconductance parameter for the load device is

$$K_L = \left(\frac{4\ \mu\text{m}}{0.5\ \mu\text{m}}\right)\left(\frac{(580\ \text{cm}^2/\text{Vs})(3.9)(8.85\times10^{-14}\ \text{F/cm})}{90\times10^{-8}\ \text{cm}}\right) = 1.78\ \text{mA}/\text{V}^2 .$$

The maximum tolerable propagation delay is about 5% of the clock period:

$$t_{P,\max} \approx 0.05T = \frac{0.05}{f} = 0.5\ \text{ns} .$$

Assuming that $t_{PLH} \gg t_{PHL}$, the maximum tolerable load capacitance is

$$C_{L,\max} = \frac{K_L V_{TL}^2 t_{P,\max}}{V_{DD}} = \frac{(1.78\times10^{-3}\ \text{A/V}^2)(-0.5\ \text{V})^2(0.5\times10^{-9}\ \text{s})}{3.3\ \text{V}} = 67fF$$

and the input capacitance for the gate is

$$C_{IN} \approx \frac{\varepsilon_{OX} W_O L_O}{t_{ox}} = \frac{(3.9)(8.85\times10^{-14}\ \text{F/cm})(2\times10^{-4}\ \text{cm})(0.5\times10^{-4}\ \text{cm})}{90\times10^{-8}\ \text{cm}} = 3.8fF .$$

Therefore,

$$N_{MAX} \le \frac{C_{L,\max}}{C_{IN}} = \frac{67fF}{3.8fF} = 17.6 ,$$

and the maximum fan-out is 17.

8.7 Logic Design

NMOS NOR gates are realized by placing switch transistors in parallel. For example, the three-input NMOS NOR3 gate is constructed as shown in Figure 8.16. In this circuit, if the voltage representing logic one is applied to any input, the associated switch MOSFET will operate in the linear region and the output will go low. If logic zero is applied to all of the inputs, then all of the switch transistors will operate in the cutoff region and the output will go high.

To a first approximation, all of the electrical properties of the NMOS NOR gate are identical to those of the NMOS inverter. Therefore, the same equations (which are approximations) can be used for hand calculations. It might

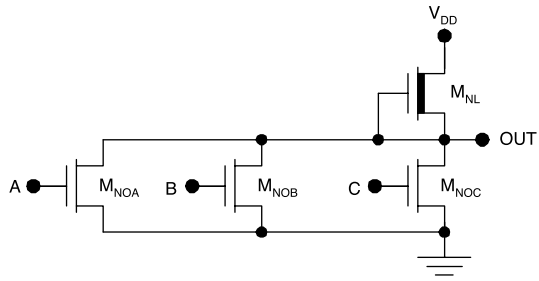


FIGURE 8.16
NMOS NOR3 gate.

appear that the high-to-low propagation delay would be reduced, compared to the inverter, by a factor of 1/M for an M-input NOR gate. After all, the M-input NOR gate can sink three times as much current as the inverter. However, this is only true if all three transistors are turned on. In the worst case, only one transistor is turned on, and the situation is identical to the inverter case if the external load capacitance is dominant. Because the paralleling of transistors does not degrade the electrical properties, it is practical to fabricate NMOS NOR gates with many inputs.

NMOS NAND gates are made by putting the switch transistors in series as shown in Figure 8.17. Most of the inverter equations also apply to the NAND gates, with some important differences. First, the aspect ratios of the switch transistors must be scaled by a factor of approximately M compared to the inverter (for an M-input NAND gate) in order to preserve the value of V_{OL}. The scaling also serves to preserve the high-to-low propagation delay.

It is not practical to implement NMOS NAND gates with a large number of inputs, for two reasons. First, scaling the switch transistors by the factor M results in a circuit with a large layout area on the chip if M is large (the chip area scales approximately with the square of M). Second, the static and

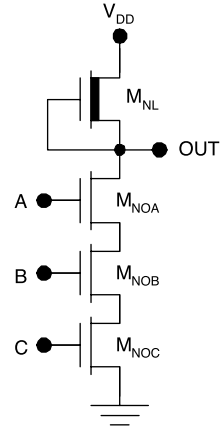


FIGURE 8.17
NMOS NAND3 gate.

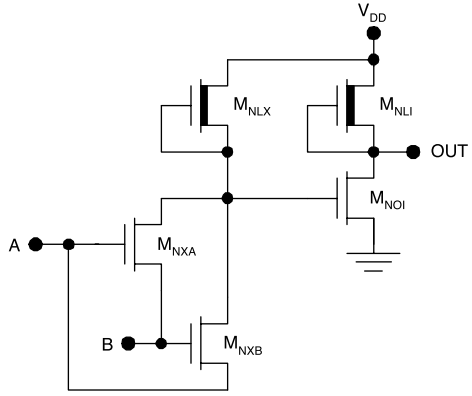


FIGURE 8.18
NMOS XOR2 gate.

dynamic performances suffer even if the switch transistors are scaled by a factor of M. This is because, under output low conditions, the drain-to-source voltage drops of the lower transistors decrease the gate-to-source biasing for the upper transistors in the M-high stack of switches. The upper transistors will therefore have a higher on-resistance and larger values of V_{DS}. The increased on-resistance will degrade the dynamic performance and the increased value of V_{DS} will degrade the voltage transfer characteristic. For these reasons, NOR gates are preferred over NAND gates, especially when the required number of inputs is large.

The exclusive OR function can be implemented in NMOS using the ingenious circuit of Figure 8.18. The operation of the circuit is as follows. If V_{INB} is logic one but V_{INA} is logic zero, M_{NXB} is linear and brings the voltage low at the gate of M_{NOI}. With M_{NOI} cut off, the output goes high. If V_{INB} is logic zero but V_{INA} is logic one, M_{NXA} is linear and brings the voltage low at the gate of M_{NOI}. With M_{NOI} cut off, the output goes high as before. On the other hand, if the inputs are the same (both logic zero or both logic one), then M_{NXA} and M_{NXB} are cut off, the voltage at the gate of M_{NOI} goes high, and the output voltage goes low.

AND–OR–INVERT gates can be implemented in NMOS by the parallel connection of series combinations of switch transistors. Figure 8.19 shows an example of this in which the inputs A, B, and C are ANDed by the series combination of the transistors M_{NOA}, M_{NOB}, and M_{NOC}. Similarly, inputs E and F are ANDed. These two results are ORed with input D by the parallel combination of the three circuit branches. The overall logic function is

$$Y = \overline{ABC + D + EF}\,. \tag{8.39}$$

Other, more complex logic functions can be realized by combining parallel and series branches of switch transistors. One such example is shown in Figure 8.20. Here, the overall logic function is

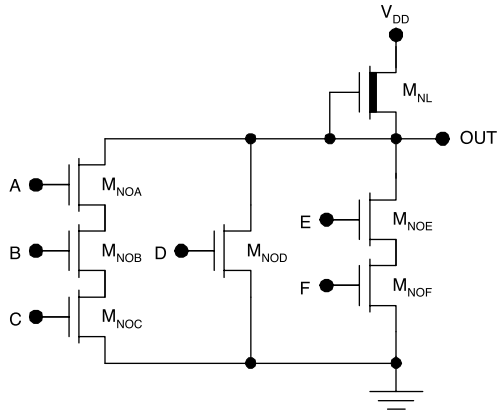


FIGURE 8.19
NMOS AND–OR–INVERT gate.

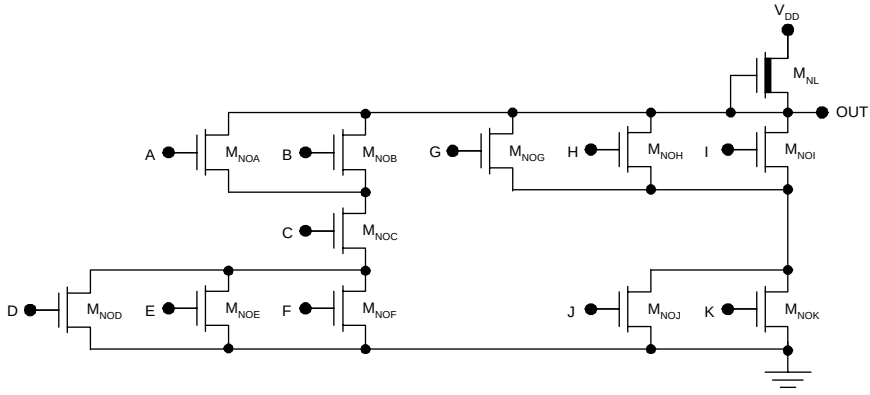


FIGURE 8.20
Implementation of a complex logic function in NMOS.

$$Y = \overline{\left[(A+B)C(D+E+G)\right]+\left[(H+I+J)(K+L)\right]}. \tag{8.40}$$

For the implementation of a general logic function in NMOS, the switch transistors in a particular branch must be scaled by a factor M compared to the transistors in an inverter. M is the maximum number of series transistors in any path from the output to ground. In the example of Figure 8.20, compared to the inverter, the switch transistors in the left branch (M_{NOA}, M_{NOB}, M_{NOC}, M_{NOD}, M_{NOE}, M_{NOF}) must be scaled by a factor of three, whereas the switch transistors in the right branch must be scaled by a factor of two (M_{NOG}, M_{NOH}, M_{NOI}, M_{NOJ}, M_{NOK}). In practice, this scaling is achieved by increasing the gate widths while keeping the gate lengths fixed.

TABLE 8.4

Depletion-Type n-MOSFET SPICE Parameters

Parameter	Value	Units
VTO	–0.3	V
KP	0.2m	A/V^2
LAMBDA	0.05	—
CGSO	1.15n	F/m
CGDO	0.58n	F/m
VMAX	100k	m/s

TABLE 8.5

Enhancement-Type n-MOSFET SPICE Parameters

Parameter	Value	Units
VTO	0.5	V
KP	0.2m	A/V^2
LAMBDA	0.05	—
CGSO	1.15n	F/m
CGDO	0.58n	F/m
VMAX	100k	m/s

8.8 PSPICE Simulations

Simulations were performed using PSPICE 9.1, which can be downloaded free.[1] The MOSFET model parameters used in all simulations are provided in Table 8.4 and Table 8.5.

The process transconductance parameters were calculated assuming an oxide thickness of 10 nm and an electron mobility of 580 cm^2/Vs:

$$KP = \frac{(3.9)\left(8.85\times10^{-14}\,\mathrm{F/cm}\right)\left(580\,\mathrm{cm^2V^{-1}s^{-1}}\right)}{10\times10^{-7}\ \mathrm{cm}} = 0.20\ \mathrm{mA/V^2}\ . \quad (8.41)$$

The oxide capacitance per unit gate width was calculated assuming a gate length of 0.5 μm and an oxide thickness of 10 nm:

$$\frac{C_{ox}}{W} = \frac{(3.9)\left(8.85\times10^{-14}\,\mathrm{F/cm}\right)\left(0.5\times10^{-4}\,\mathrm{cm}\right)}{10\times10^{-7}\,\mathrm{cm}} = 17.3\ \mathrm{pF/cm} = 1.73\ \mathrm{nF/m} \quad (8.42)$$

The individual contributions were estimated using

$$CGSO \approx \left(\frac{2}{3}\right)\left(\frac{C_{ox}}{W}\right) = 1.15 \text{ nF/m} \tag{8.43}$$

and

$$CGDO \approx \left(\frac{1}{3}\right)\left(\frac{C_{ox}}{W}\right) = 0.58 \text{ nF/m} . \tag{8.44}$$

8.8.1 Voltage Transfer Characteristic

The voltage transfer characteristic was simulated for the NMOS inverter of Figure 8.21 with V_{DD} = 2.5 V; the results are shown in Figure 8.22. The critical voltages are: V_{OH} = 5 V, V_{OL} = 0.056 V, V_{IL} = 0.61 V, and V_{IH} = 1.26 V.

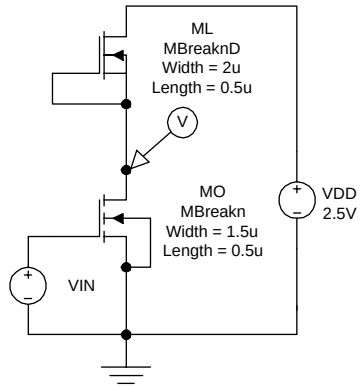


FIGURE 8.21
NMOS circuit used for simulation of the VTC.

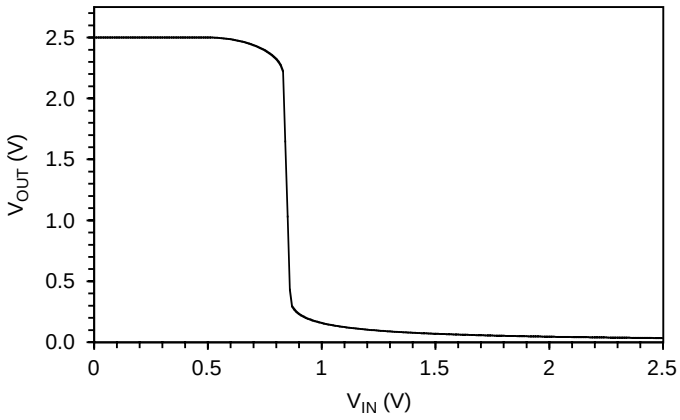


FIGURE 8.22
Simulated VTC for the NMOS circuit.

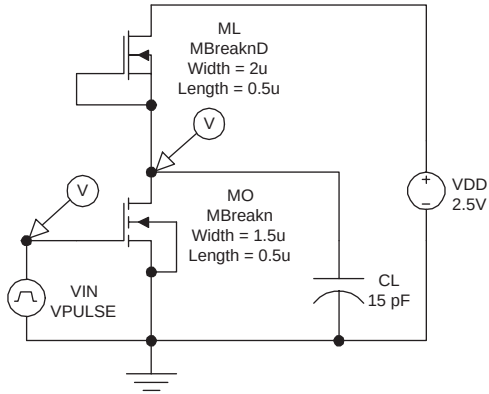


FIGURE 8.23
NMOS circuit used for the simulation of the propagation delays.

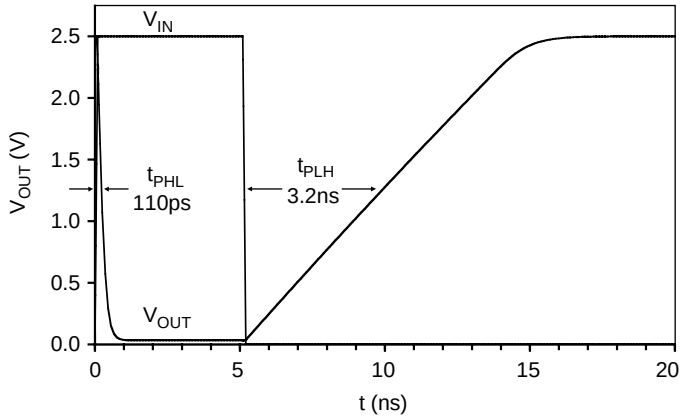


FIGURE 8.24
Simulated NMOS transient response.

8.8.2 Propagation Delays

The propagation delays for the NMOS inverter with a 15-pF load were determined using the circuit of Figure 8.23. The pulse source parameters were V1 = 0 V, V2 = 2.5, TD = TR = TF = 0, PW = 5 ns, and PER = 20 ns. The results of the transient simulation appear in Figure 8.24.

The propagation delays, t_{PHL} = 110 ps and t_{PLH} = 3.2 ns, can be determined using the 50% points on the input and output waveforms. Additional transient simulations were done to determine the propagation delays as a function of the load capacitance and the results are shown in Figure 8.25. Both propagation delays increased linearly with the load capacitance. The slopes of the characteristics are 10 Ω (t_{PHL}) and 310 Ω (t_{PLH}). A log scale was used to show both characteristics on the same graph.

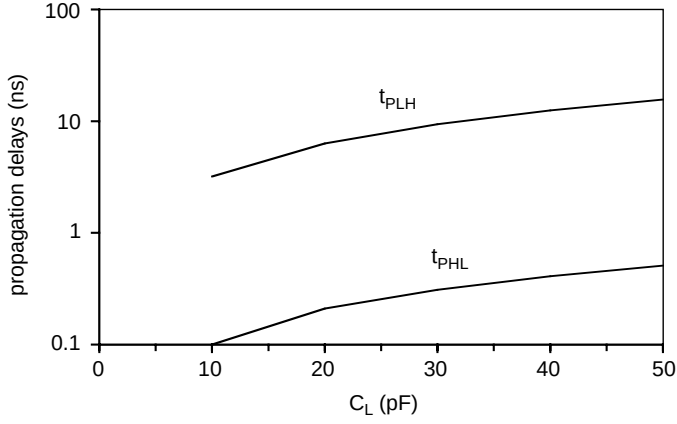


FIGURE 8.25
NMOS propagation delays as a function of load capacitance.

8.9 Summary

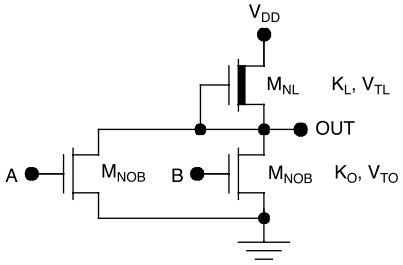
NMOS logic gates are constructed using only n-channel MOSFETs. Of the several implementations of NMOS logic, the most common scheme involves using a depletion-type n-MOSFET pull-up device.

NMOS has the highest packing density of any logic family. The primary drawback of NMOS circuits is their appreciable DC dissipation. CMOS has much lower DC dissipation and has replaced NMOS in most VLSI applications, although NMOS finds use in high-density memories. Also, complementary MOSFETs have been used to realize some alternative logic circuits similar to NMOS.

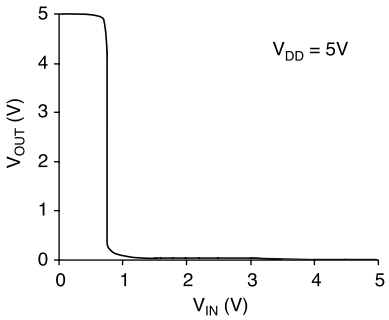
Complex logic functions can be realized in NMOS by combining series and parallel branches of n-channel MOSFETs. Placing n-MOSFETs in series provides the ANDing of inputs while the paralleling of n-MOSFETs provides the OR function. In such complex gates, the MOSFETs must be scaled appropriately to maintain acceptable DC and transient characteristics.

NMOS LOGIC QUICK REFERENCE

NMOS Inverter Circuit



DC Voltage Transfer Characteristic



NMOS logic circuits use only n-channel MOSFETs, resulting in the most efficient utilization of chip area. The primary drawback of NMOS is its high static dissipation compared to CMOS.

DC Voltage Transfer Characteristic

$$V_{OH} = V_{DD} \qquad V_{OL} = V_{DD} - V_{TO} \pm \sqrt{(V_{DD} - V_{TO})^2 - \left(\frac{K_L}{K_O}\right) V_{TL}^2}$$

$$V_{IL} = V_{TO} + \frac{K_L}{\sqrt{K_O K_L + K_O^2}} |V_{TL}| \qquad V_{IH} = V_{TO} + 2|V_{TL}| \sqrt{\frac{K_L}{3K_O}}$$

Dissipation

$$P = P_{DC} + P_{AC} = \frac{K_L V_{DD} V_{TL}^2}{4} + f C_L V_{DD}^2$$

Propagation Delays

$$t_{PLH} = \frac{V_{DD} C_L}{K_L V_{TL}^2} \qquad t_{PHL} \approx \frac{V_{DD} C_L}{K_O (V_{DD} - V_{TO})^2}$$

Fan-out

$$N_{MAX} \le \frac{K_L V_{TL}^2 t_{P,\max}}{V_{DD} C_{IN}} \qquad C_{IN} \approx \frac{\varepsilon_{OX} W_O L_O}{t_{OX}}$$

Design Rules

$$\mu_n = 580 cm^2 V^{-1} s^{-1} \qquad \varepsilon_{OX} = 3.9\varepsilon_0$$

f = 10^{-15} p = 10^{-12} n = 10^{-9} μ = 10^{-6} m = 10^{-3} k = 10^{3} M = 10^{6} G = 10^{9}

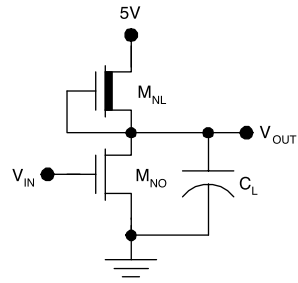


FIGURE 8.26
Circuit diagram (L8.1).

Laboratory Exercises

L8.1. Obtain commercially available enhancement-type and depletion-type n-channel MOSFETs (seven of each) for the construction of depletion-loaded NMOS inverters as shown in Figure 8.26. From the manufacturer's Web page, obtain the data sheets for both types of transistors.

1. Determine and plot the unloaded voltage transfer characteristic using hand calculations.
2. Determine and plot the unloaded voltage transfer characteristic using SPICE.
3. Estimate P_H, P_L, and P_{DC} by hand calculations.
4. Determine P_H, P_L, and P_{DC} using SPICE.
5. Using hand calculations, estimate the load capacitance for which the average propagation delay will be 1000 ns.
6. Using SPICE, estimate the propagation delays for the load capacitance determined in part (5).
7. Build the inverter and measure the unloaded voltage transfer characteristic using the x–y feature of an oscilloscope or virtual instrument and a low-frequency input signal (1 kHz).
8. Build a seven-stage ring oscillator using seven of these inverter circuits. Load each stage with the capacitance determined in part (5). Measure the frequency of oscillation and determine the average propagation delay from the measured frequency of oscillation.

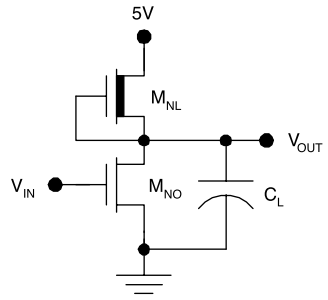


FIGURE 8.27
Circuit diagram (L8.2).

L8.2. Obtain commercially available enhancement-type and depletion-type n-channel MOSFETs (seven of each) for the construction of depletion-loaded NMOS inverters as illustrated in Figure 8.27. From the manufacturer's Web page, obtain the data sheets for both types of transistors.

1. Using hand calculations, estimate and plot the average propagation delay vs. the load capacitance.
2. Using SPICE, calculate and plot the average propagation delay vs. the load capacitance.
3. Build a seven-stage ring oscillator using seven of these inverter circuits. Determine and plot the average propagation delay vs. the load capacitance.
4. Is there a range of C_L for which t_P is independent of the load capacitance? Is this predicted by SPICE? Is this predicted by hand calculations? Explain.
5. What is the range of C_L for which $t_P \propto C_L$?

L8.3. Obtain commercially available enhancement-type and depletion-type n-channel MOSFETs (seven of each) for the construction of depletion-loaded NMOS inverters as shown in Figure 8.28. From the manufacturer's Web page, obtain the data sheets for both types of transistors.

1. Using hand calculations, estimate the value of C_L for which the average propagation delay is 1000 ns. Use this value of load capacitance for the remaining parts.
2. Using hand calculations, estimate and plot the average propagation delay vs. the supply voltage, the average quiescent dissipation vs. the supply voltage, and the power delay product vs. the supply voltage.

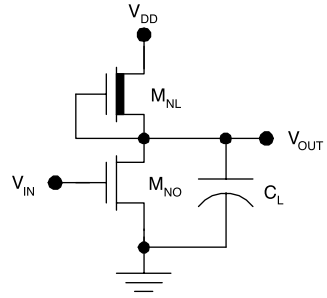


FIGURE 8.28
Circuit diagram (L8.3).

3. Using SPICE, calculate and plot the average propagation delay vs. the supply voltage, the average quiescent dissipation vs. the supply voltage, and the power delay product vs. the supply voltage.
4. Build a single NMOS inverter and make measurements to determine the average quiescent dissipation.
5. Build a seven-stage ring oscillator using seven of these inverter circuits. Based on the experimental results, plot the average propagation delay vs. the supply voltage, the average quiescent power vs. the supply voltage, and the power delay product vs. the supply voltage.
6. Based on results, what recommendation should be made for the choice of the supply voltage?

L8.4. Obtain commercially available enhancement-type and depletion-type n-channel MOSFETs (seven of each) for the construction of depletion-loaded NMOS inverters as shown in the circuit diagram of Figure 8.29. From the manufacturer's Web page, obtain the data sheets for both types of transistors.

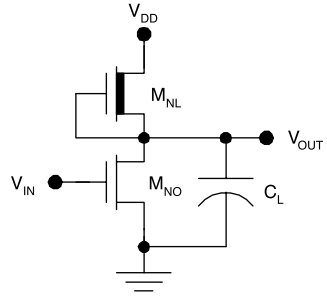


FIGURE 8.29
Circuit diagram (L8.4).

1. Using hand calculations, estimate and plot the average propagation delay vs. the load capacitance, with the supply voltage as a parameter.
2. Using SPICE, estimate and plot the average propagation delay vs. the load capacitance, with the supply voltage as a parameter.
3. Build a seven-stage ring oscillator using seven of these inverter circuits. Based on the experimental results, plot the average propagation delay vs. the load capacitance, with the supply voltage as a parameter.
4. Is it possible to fit the results with an expression of the form

$$t_P = A\frac{C_L}{V_{DD}},$$

where A is a constant?

Problems

P8.1. Consider the NMOS gate depicted in Figure 8.30.

1. Determine V_{IL}, V_{IH}, V_{OL}, and V_{OH}.
2. Determine the noise margins V_{NML} and V_{NMH}.
3. Estimate the average DC dissipation.

P8.2. Consider the NMOS gate of Figure 8.31. Determine the values of V_{IN} and V_{OUT} for which M_{NO} is at the boundary between saturation and ohmic operation.

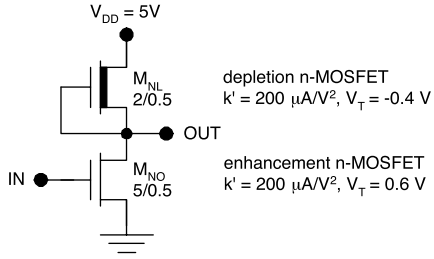


FIGURE 8.30
NMOS gate (P8.1).

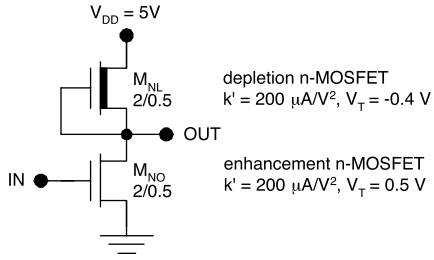


FIGURE 8.31
NMOS gate (P8.2).

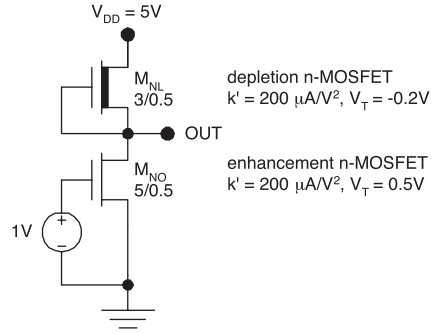


FIGURE 8.32
NMOS gate (P8.3).

P8.3. Consider the NMOS gate illustrated in Figure 8.32.
1. Determine the mode of operation for each of the transistors.
2. Determine the supply current I_{DD}.
3. Determine the value of V_{OUT}.

P8.4. Consider the NMOS gate of Figure 8.33.
1. Determine the mode of operation for each of the transistors.
2. Determine the supply current I_{DD}.
3. Determine the value of V_{OUT}.

P8.5. Consider the NMOS gate illustrated in Figure 8.34.
1. Using hand calculations, determine and plot the voltage transfer characteristic for the inverter.
2. Determine the range of V_{IN} for each mode of operation (cutoff, saturation, and linear) of M_{NO}.

P8.6. Consider the NMOS gate shown in Figure 8.35.
1. Determine the value of V_{OUT} if V_A = 5 V and $V_B = V_C$ = 0 V. (This is the worst case of V_{OL}.)
2. Determine the value of V_{OUT} if $V_A = V_B = V_C$ = 5 V.

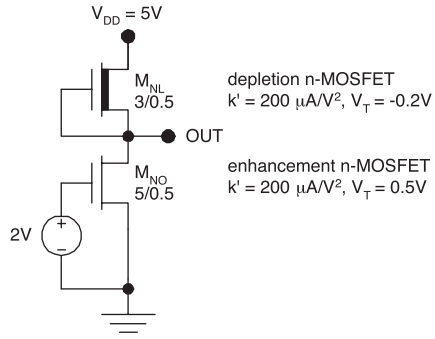


FIGURE 8.33
NMOS gate (P8.4).

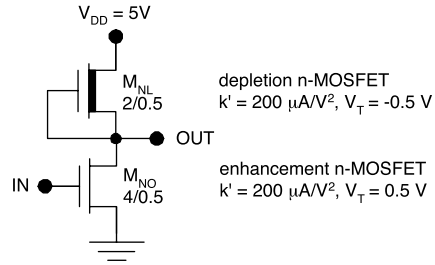


FIGURE 8.34
NMOS gate (P8.5).

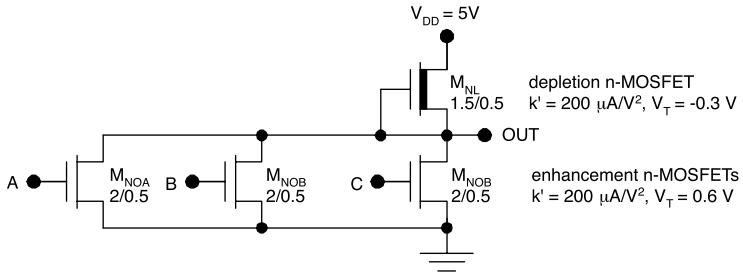


FIGURE 8.35
NMOS gate (P8.6).

P8.7. Consider the NMOS NAND3 gate illustrated in Figure 8.36. Determine the value of V_{OL}.

P8.8. Consider the NMOS gate depicted in Figure 8.37.

1. Using hand calculations, determine V_{IL}, V_{IH}, V_{OL}, and V_{OH}.
2. Using SPICE, determine and plot the voltage transfer characteristic for the gate.
3. From the tabulated SPICE results, determine the critical voltages and compare these to the values found in (a).

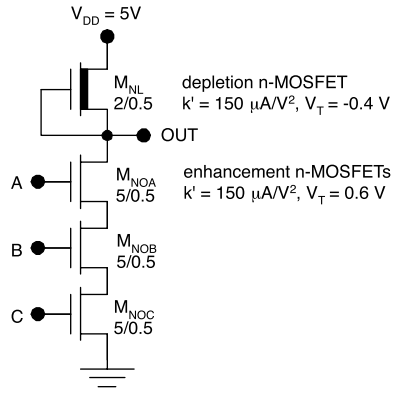


FIGURE 8.36
NMOS NAND3 gate (P8.7).

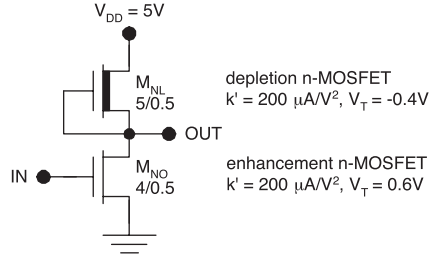


FIGURE 8.37
NMOS gate (P8.8).

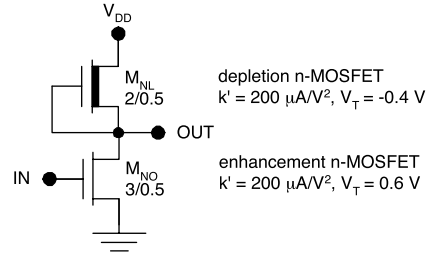


FIGURE 8.38
NMOS gate (P8.9).

P8.9. Consider the NMOS gate of Figure 8.38.

1. Using hand calculations, calculate and plot the DC power dissipation vs. the supply voltage.
2. Repeat using SPICE. Plot the results together on one graph for comparison.

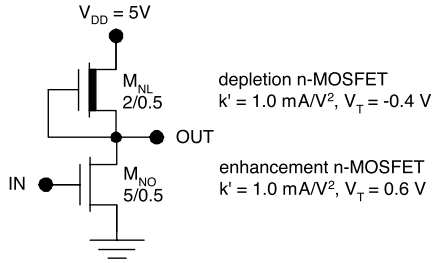


FIGURE 8.39
NMOS gate (P8.10).

P8.10. Consider the NMOS gate shown in Figure 8.39.

1. From the information given, estimate the oxide thickness used in the transistors.
2. Calculate the approximate input capacitance for the inverter.
3. Estimate the maximum fan-out if $t_{P,\max} = 10$ ns.

P8.11. Consider depletion-loaded NMOS with $V_{DD} = 5$ V, $V_{TO} = 0.5$ V, $V_{TL} = -0.6$ V, and $t_{ox} = 18$ nm. All devices have 0.5 μm gate lengths. Choose the W_O/W_L ratio such that $V_{OL} = 25$ mV.

P8.12. Consider depletion-loaded NMOS with $V_{DD} = 3.3$ V, $V_{TO} = 0.5$ V, $V_{TL} = -0.3$ V, and $t_{ox} = 10$ nm. All devices have 0.35 μm gate lengths. Design the transistors for the inverter (choose W_O and W_L) such that $V_{OL} = 20$ mV and the average DC dissipation is 2 mW.

P8.13. Consider depletion-loaded NMOS with $V_{DD} = 3.3$ V, $V_{TO} = 0.5$ V, $V_{TL} = -0.3$ V, and $t_{ox} = 10$ nm. All devices have 0.35 μm gate lengths. Design the transistors for the inverter (choose W_O and W_L) such that $V_{OL} = 20$ mV and $t_P \leq 10$ ns with $C_L = 15$ pF. (Consider the worst-case propagation delay.)

P8.14. Consider depletion-loaded NMOS with $V_{DD} = 3.3$ V, $V_{TO} = 0.5$ V, $V_{TL} = -0.3$V, and $t_{ox} = 9$ nm. All devices have 0.3 μm gate lengths. Is it possible to design the transistors for the inverter such that $V_{OL} = 20$ mV, $t_{PLH} \leq 10$ ns with $C_L = 15$ pF and $P_{DC} \leq 1$ mW? Provide a design satisfying these requirements or show why one does not exist.

Reference

1. www.cadence.com (Cadence).

9

CMOS Logic

9.1 Introduction

Complementary metal oxide–semiconductor (CMOS) logic is by far the most important logic family today. CMOS takes its name from the fact that the circuits use equal numbers of complementary n-channel and p-channel MOSFETs. The overwhelming popularity of CMOS has come about due to its extremely low standby power dissipation, high packing density, and high speed. CMOS is used extensively in high-performance, portable (battery-operated) products such as notebook computers, portable digital assistants, pagers, and wireless phones.

A CMOS inverter comprises one enhancement type n-MOSFET and one enhancement type p-MOSFET as shown in Figure 9.1. Normally-off devices are used for low standby dissipation. With a logic-zero input ($V_{IN} = 0$), the n-MOSFET is cut off but the p-MOSFET is linear. Therefore, the output goes high to V_{DD}. With a logic-one input ($V_{IN} = V_{DD}$), the n-MOSFET is linear but the p-MOSFET is cut off, so the output goes low to 0. In either case, the only contribution to the supply current is the small leakage current in a cutoff MOSFET; for this reason DC dissipation is negligible. This unique property of CMOS has made it the logic family of choice for battery-operated products.

The packing density of CMOS is very high because the circuits use only MOSFETs. MOSFETs are 1/10 the size of bipolar transistors and 1/500 the size of resistors in terms of chip area. Only NMOS provides a higher packing density than CMOS. NMOS has this advantage because it requires only n-MOSFETs, so it is not necessary to fabricate p-type and n-type wells. However, for most applications, the slight packing density advantage of NMOS is overshadowed by the superior DC dissipation of CMOS.

In terms of speed, CMOS has made rapid gains in recent years. CMOS gates now exhibit on-chip propagation delays in picoseconds. In silicon technology, only ECL offers greater speed in this regard. However, the off-chip data rates for CMOS are markedly slower than those for bipolar logic gates. This is the one area in which bipolar logic circuits still hold an advantage

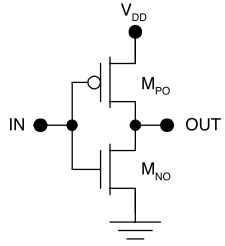


FIGURE 9.1
Complementary metal oxide–semiconductor (CMOS) inverter.

over CMOS. As this gap diminishes, CMOS will probably gain additional market share in the industry it already dominates.

9.2 Voltage Transfer Characteristic

Consider the voltage transfer characteristic for the CMOS inverter shown in Figure 9.2. To determine the voltage transfer characteristic, five regimes of operation must be considered. These will be discussed next in order of increasing input voltage.

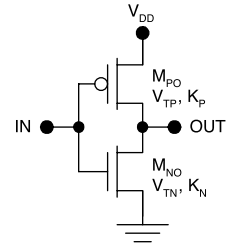


FIGURE 9.2
CMOS inverter.

9.2.1 n-MOSFET Cutoff, p-MOSFET Linear

If the input voltage is less than the threshold voltage of the n-MOSFET, then the n-MOSFET is cut off but the p-MOSFET is linear. For this situation, the output voltage is equal to V_{DD}:

$$V_{OH} = V_{DD}. \tag{9.1}$$

9.2.2 n-MOSFET Saturated, p-MOSFET Linear

If the input voltage is increased somewhat beyond the threshold voltage of the n-MOSFET, the n-MOSFET will be saturated while the p-MOSFET will remain linear. The condition for saturation operation of M_{NO} is

$$\left(V_{\text{IN}} \geq V_{TN}\right) \text{ and } \left(V_{\text{IN}} - V_{TN} \leq V_{\text{OUT}}\right). \tag{9.2}$$

If these conditions are satisfied, then the supply current is equal to the saturated drain current for the n-channel device:

$$I_{DD} = \frac{K_N\left(V_{IN} - V_{TN}\right)^2}{2}. \tag{9.3}$$

The output voltage can be determined using the drain-to-source voltage for the linear device:

$$V_{\text{OUT}} = V_{DD} + V_{DSPO}, \tag{9.4}$$

where V_{DSPO} is the (negative) drain-to-source voltage for the p-MOSFET and

$$\begin{aligned} V_{DSPO} &= (V_{GSPO} - V_{TP}) + \sqrt{\left(V_{GSPO} - V_{TP}\right)^2 - \frac{2I_{DD}}{K_P}} \\ &= (V_{\text{IN}} - V_{DD} - V_{TP}) + \sqrt{\left(V_{\text{IN}} - V_{DD} - V_{TP}\right)^2 - \frac{K_N}{K_P}\left(V_{\text{IN}} - V_{TN}\right)^2} \end{aligned} \tag{9.5}$$

Therefore,

$$V_{\text{OUT}} = (V_{\text{IN}} - V_{TP}) + \sqrt{\left(V_{\text{IN}} - V_{DD} - V_{TP}\right)^2 - \frac{K_N}{K_P}\left(V_{\text{IN}} - V_{TN}\right)^2}. \tag{9.6}$$

It must be emphasized that this equation only applies for the situation in which the n-MOSFET is saturated and the p-MOSFET is linear (Equation 9.2). Because V_{OUT} is not known *a priori*, Equation 9.6 should be used to calculate the output voltage and then the conditions in Equation 9.2 should be checked to verify that the results are meaningful.

9.2.3 Both MOSFETs Saturated

If V_{IN} is further increased, the p-MOSFET will become saturated. Therefore, both MOSFETs are saturated if

$$\left(V_{\text{IN}} - V_{TN} \le V_{\text{OUT}}\right) \text{ and } \left(V_{\text{IN}} - V_{TP} \ge V_{\text{OUT}}\right). \tag{9.7}$$

With both MOSFETs saturated, the voltage transfer characteristic cannot be calculated without knowledge of the channel length modulation parameters. For hand calculations, the best approach is to interpolate between the two adjacent regions of the voltage transfer characteristic.

9.2.4 n-MOSFET Linear, p-MOSFET Saturated

With the n-MOSFET linear and the p-MOSFET saturated, the supply current is equal to the drain current in the p-MOSFET and the output voltage is equal to the drain-to-source voltage for the n-MOSFET. The conditions for linear operation of the n-MOSFET and saturated operation of the p-MOSFET are

$$\left(V_{\text{IN}} - V_{TN} \geq V_{\text{OUT}}\right) \text{ and } \left(V_{\text{IN}} \leq V_{DD} + V_{TP}\right), \tag{9.8}$$

respectively. Under these conditions (which must be verified after the calculation of V_{OUT}),

$$I_{DD} = \frac{K_P\left(V_{\text{IN}} - V_{DD} - V_{TP}\right)^2}{2} \tag{9.9}$$

and

$$V_{\text{OUT}} = (V_{\text{IN}} - V_{TN}) + \sqrt{\left(V_{\text{IN}} - V_{TN}\right)^2 - \frac{K_P}{K_N}\left(V_{\text{IN}} - V_{DD} - V_{TP}\right)^2}\,. \tag{9.10}$$

9.2.5 n-MOSFET Linear, p-MOSFET Cutoff

If the input voltage is sufficiently close to V_{DD}, then the p-MOSFET will cut off. This occurs if

$$\left(V_{\text{IN}} \geq V_{DD} + V_{TP}\right). \tag{9.11}$$

In this case, zero supply current flows and the output goes to zero. Thus

$$V_{OL} = 0\,. \tag{9.12}$$

The logic swing of CMOS is therefore equal to the supply voltage. This is called "rail-to-rail" operation and CMOS is said to have a rail-to-rail logic swing — a unique advantage of CMOS.

9.2.6 Summary of Voltage Transfer Characteristic

In summary, the voltage transfer characteristic for a CMOS inverter can be calculated in five pieces, corresponding to the different modes of operation for the two devices. These results are summarized in Table 9.1 for a symmetric CMOS inverter, for which $K_N = K_P = K$ and $V_{TN} = |V_{TP}| = V_T$. Notice

TABLE 9.1

Equations for the Voltage Transfer Characteristic of the CMOS Inverter with Matched MOSFETs[a]

	Range of V_{IN}	p-MOS mode	n-MOS mode	Equation for V_{OUT}
1	$V_{IN} \le V_T$	Linear	Cutoff	$V_{OUT} = V_{DD}$
2	$V_T \le V_{IN} \le V_{OUT} - V_T$	Linear	Saturation	$V_{OUT} = (V_{IN} + V_T) + \sqrt{(V_{IN} - V_{DD} + V_T)^2 - (V_{IN} - V_T)^2}$
3	$V_{OUT} - V_T \le V_{IN} \le V_{OUT} + V_T$	Saturation	Saturation	Interpolate
4	$V_{OUT} + V_T \le V_{IN} \le V_{DD} - V_T$	Saturation	Linear	$V_{OUT} = (V_{IN} - V_T) - \sqrt{(V_{IN} - V_T)^2 - (V_{IN} - V_{DD} + V_T)^2}$
5	$V_{IN} \ge V_{DD} - V_T$	Cutoff	Linear	$V_{OUT} = 0$

[a] $K_N = K_P = K$ and $V_{TN} = |V_{TP}| = V_T$.

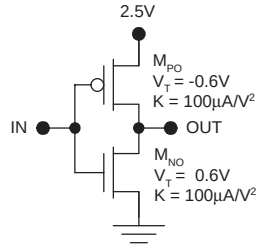


FIGURE 9.3
Example CMOS inverter for calculation of the voltage transfer characteristic.

that the voltage transfer characteristic has an odd symmetry about the midpoint where $V_{IN} = V_{OUT} = V_{DD}/2$.

Example 9.1

Calculate the voltage transfer characteristic for a symmetric CMOS inverter of Figure 9.3 with $V_{DD} = 2.5$ V, $V_T = 0.6$ V, and $K = 100\ \mu A/V^2$.

Solution. The voltage transfer characteristic can be calculated piecewise as follows.

$$V_{OUT} = \begin{cases} 2.5\text{ V}; & V_{IN} \le 0.6\text{ V} \\ V_{IN} + 0.6\text{ V} + \sqrt{\left(V_{IN} - 1.9\text{ V}\right)^2 - \left(V_{IN} - 0.6\text{ V}\right)^2}; & 0.6\text{ V} \le V_{IN} \le 1.25\text{ V} \\ 1.25V; & V_{IN} = 1.25\text{ V} \\ V_{IN} - 0.6\text{ V} - \sqrt{\left(V_{IN} - 0.6\text{ V}\right)^2 - \left(V_{IN} - 1.9\text{ V}\right)^2}; & 1.25\text{ V} \le V_{IN} \le 1.9\text{ V} \\ 0; & V_{IN} \ge 1.9\text{ V} \end{cases}$$

For this example, both transistors are saturated only at $V_{IN} = V_{DD}/2$, so no interpolation is needed. The characteristic is plotted in Figure 9.4. For the symmetric CMOS inverter with matched transistors, the characteristic exhibits odd symmetry about the $V_{DD}/2$ point.

9.3 Short-Circuit Current in CMOS

In a CMOS gate, normally one of the two MOSFETs cuts off. With a logic-one input, the p-MOSFET is cut off and, with a logic-zero input, the n-MOSFET is cut off. In either situation there is no path between V_{DD} and ground for DC current. However, significant supply current can flow if the input voltage is between zero and V_{DD}.

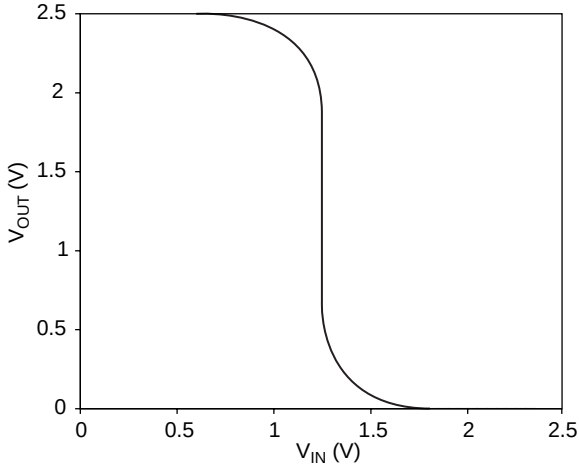


FIGURE 9.4
Example voltage transfer characteristic for a symmetric CMOS inverter with V_{DD} = 2.5 V, V_T = 0.6 V, and K = 100 μA/V².

Consider a symmetric CMOS inverter and suppose the input voltage is gradually increased from zero to V_{DD}. If the input voltage is less than V_T, the n-MOSFET will be cut off and no current will flow:

$$I_{DD} = 0; \qquad \left(V_{\text{IN}} \le V_T\right). \tag{9.13}$$

As shown in the previous section, the n-MOSFET is saturated if the input voltage is between V_T and one half of the supply voltage. Therefore,

$$I_{DD} = \frac{K\left(V_{\text{IN}} - V_T\right)^2}{2}; \qquad \left(V_T \le V_{\text{IN}} \le \frac{V_{DD}}{2}\right). \tag{9.14}$$

Similarly, the p-MOSFET is saturated if the input voltage is between $V_{DD}/2$ and $V_{DD} - V_T$. Thus,

$$I_{DD} = \frac{K\left(V_{\text{IN}} - V_{DD} + V_T\right)^2}{2}; \qquad \left(\frac{V_{DD}}{2} \le V_{\text{IN}} \le V_{DD} - V_T\right). \tag{9.15}$$

Finally, the p-MOSFET is cut off for more positive values of the input voltage:

$$I_{DD} = 0; \qquad \left(V_{\text{IN}} \ge V_{DD} - V_T\right). \tag{9.16}$$

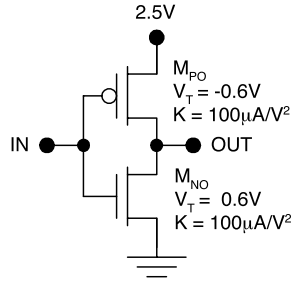


FIGURE 9.5
Example CMOS inverter for calculation of the short-circuit current vs. the input voltage.

Example 9.2

Calculate the short-circuit current vs. the input voltage for the symmetric CMOS inverter illustrated in Figure 9.5 with V_{DD} = 2.5 V, V_T = 0.6 V, and K = 100 μA/V².

Solution. The short circuit current vs. the input voltage is given by

$$I_{DD} = \begin{cases} 0; & V_{\text{IN}} \le 0.6 \text{ V} \\ 50 \ \mu\text{A/V}^2 \left(V_{\text{IN}} - 0.6 \text{ V}\right)^2; & 0.6 \text{ V} \le V_{\text{IN}} \le 1.25 \text{ V} \\ 50 \ \mu\text{A/V}^2 \left(V_{\text{IN}} - 1.9 \text{ V}\right)^2; & 1.25 \text{ V} \le V_{\text{IN}} \le 1.9 \text{ V} \\ 0; & V_{\text{IN}} \ge 1.9 \text{ V} \end{cases}$$

The results are plotted in Figure 9.6. For the symmetric CMOS inverter with matched transistors, the characteristic is symmetric and the peak current flows when the input voltage is equal to one half of the supply voltage.

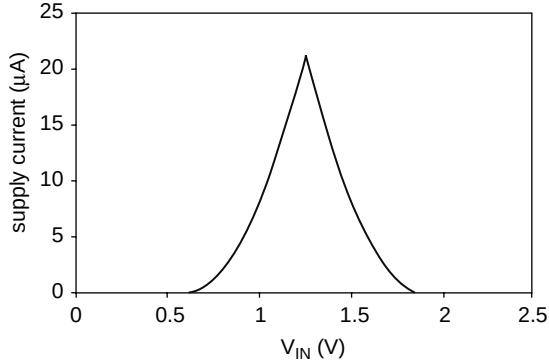


FIGURE 9.6
Short-circuit current vs. input voltage characteristic for a symmetric CMOS inverter with V_{DD} = 2.5 V, V_T = 0.6 V, and K = 100 μA/V².

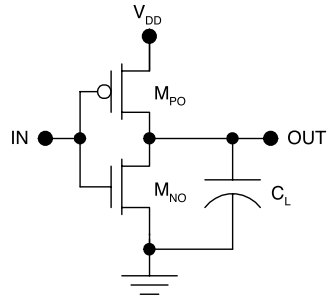


FIGURE 9.7
Symmetric CMOS inverter with a lumped capacitive load.

9.4 Propagation Delays

Consider a symmetric CMOS inverter with a lumped capacitive load as shown in Figure 9.7. For the purpose of estimating the propagation delays, assume abrupt voltage transitions at the input. First, consider the high-to-low propagation delay, t_{PHL}. Suppose the input voltage makes an abrupt transition from zero to V_{DD} at $t = 0$. The n-MOSFET becomes saturated at $t = 0^+$ and remains in this mode of operation until V_{OUT} drops to $V_{DD} - V_T$. During this time interval, a constant current flows in M_{NO}:

$$I_{DNO} = \frac{K\left(V_{DD} - V_T\right)^2}{2}. \tag{9.17}$$

The n-MOSFET becomes linear when $V_{\text{OUT}} = V_{DD} - V_T$ at $t = T_{D1}$:

$$T_{D1} = \frac{2V_T C_L}{K\left(V_{DD} - V_T\right)^2}. \tag{9.18}$$

Once V_{OUT} drops below $V_{DD} - V_T$, the n-MOSFET moves into the linear region of operation. As a first approximation for the linear n-MOSFET,

$$I_{DN} = \frac{V_{\text{OUT}}}{R_{DN}}, \tag{9.19}$$

where

$$\frac{1}{R_{DN}} = \text{average value of} \left(\frac{\partial I_{DN}}{\partial V_{\text{OUT}}} \right). \tag{9.20}$$

For the linear n-MOSFET, the drain current is given by

$$I_{DN} = K\left[\left(V_{DD} - V_T\right)V_{\text{OUT}} - \frac{V_{\text{OUT}}^2}{2}\right]. \tag{9.21}$$

The partial derivative of the drain current with respect to the output voltage is

$$\frac{\partial I_{DN}}{\partial V_{\text{OUT}}} = K\left[\left(V_{DD} - V_T\right) - V_{\text{OUT}}\right]. \tag{9.22}$$

The average value of the partial derivative may be estimated by substituting in the average value of V_{OUT}, yielding

$$\frac{1}{R_{DN}} = \frac{K\left(V_{DD} - V_T\right)}{2}, \tag{9.23}$$

By using the ohmic approximation for the linear n-MOSFET, the circuit can be treated as a simple *RC* combination. Thus,

$$V_{\text{OUT}} \approx \left(V_{DD} - V_T\right)\exp\left[-\frac{\left(t - T_{D1}\right)}{R_{DN}C_L}\right]. \tag{9.24}$$

The output voltage reaches the 50% point at $t = T_{D1} + T_{D2}$, where

$$T_{D2} = R_{DN}C_L \ln\left(\frac{V_{DD} - V_T}{V_{DD}/2}\right). \tag{9.25}$$

The high-to-low propagation delay is the sum of T_{D1} and T_{D2}:

$$t_{PHL} = T_{D1} + T_{D2} \approx \frac{C_L}{K}\left[\frac{2V_T}{\left(V_{DD} - V_T\right)^2} + \frac{2}{\left(V_{DD} - V_T\right)}\ln\left(\frac{V_{DD} - V_T}{V_{DD}/2}\right)\right]. \tag{9.26}$$

The analysis of t_{PLH} is very similar. In fact, for a symmetric CMOS inverter, the two propagation delays are equal. Therefore,

$$t_{PLH} = t_{PHL} = t_P, \tag{9.27}$$

where

$$t_P \approx \frac{C_L}{K}\left[\frac{2V_T}{(V_{DD}-V_T)^2}+\frac{2}{(V_{DD}-V_T)}\ln\left(\frac{V_{DD}-V_T}{V_{DD}/2}\right)\right]. \tag{9.28}$$

This result shows that the propagation delays are proportional to the load capacitance and inversely proportional to the device transconductance parameters. Higher speed can be obtained by reducing the load capacitance or by scaling up the K values. In addition, there is an approximate inverse dependence on the supply voltage, which has led to occasional use of the bold approximation

$$t_P \approx \frac{2C_L}{KV_{DD}}. \tag{9.29}$$

This bold approximation is surprisingly useful; it typically yields errors less than 25% for high-voltage CMOS circuits. In low-voltage circuits, another approximation is sometimes used to take into account the dependence on the threshold voltage:

$$t_P \approx \frac{1.6C_L}{K(V_{DD}-V_T)}. \tag{9.30}$$

Example 9.3

Estimate t_P (15-pF load) for the symmetric CMOS inverter depicted in Figure 9.8 with V_{DD} = 2.5 V, V_T = 0.6 V, and K = 0.5 mA/V².

Solution. The propagation delay is

$$t_P \approx \frac{15\times10^{-12}\ \text{F}}{0.5\times10^{-3}\ \text{A/V}^2}\left[\frac{1.2\ \text{V}}{(1.9\ \text{V})^2}+\frac{2}{1.9\ \text{V}}\ln\left(\frac{1.9\ \text{V}}{1.25\ \text{V}}\right)\right]=23\ \text{ns}.$$

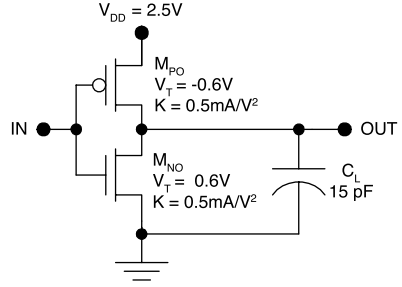


FIGURE 9.8
Example of CMOS inverter with a 15-pF load.

9.5 Dissipation

Broadly speaking, the dissipation in CMOS can be classified as static (DC) or dynamic (AC) dissipation. The total dissipation is the sum of these two components:

$$P = P_{DC} + P_{AC} . \tag{9.31}$$

Static dissipation is associated with the subthreshold currents in the cutoff MOSFETs and the leakage currents in the reverse-biased source and drain p–n junction diodes. The dynamic dissipation is the sum of the short-circuit power and the capacitance switching power. Thus,

$$P = \underbrace{P_{subthreshold} + P_{pn}}_{P_{DC}} + \underbrace{P_{sc} + P_{switch}}_{P_{AC}} , \tag{9.32}$$

where

$P_{subthreshold}$ = power associated with MOSFET subthreshold conduction
P_{pn} = power associated with p–n junction leakage in the MOSFETs
P_{SC} = short-circuit dissipation
P_{switch} = capacitance switching dissipation

Usually, the capacitance switching power is dominant in CMOS circuitry. However, all four components must be considered in low-power VLSI design.

9.5.1 Capacitance Switching Dissipation

Suppose a CMOS gate is loaded by a lumped capacitance, as shown in Figure 9.9, and the output of this gate switches from low to high and back to low again. The energy drawn from the power supply during the low-to-high transition is

$$J_{switch} = V_{DD} C_L \int_0^{V_{DD}} dV = C_L V_{DD}^2 . \tag{9.33}$$

Half of this energy is stored in the capacitor and the other half is dissipated in the p-MOSFET. During the high-to-low transition, the energy stored in the capacitor is dissipated in the n-MOSFET, but no additional energy is drawn from the power supply. Thus, if the output is switched at a frequency, f, the power dissipation is

$$P_{switch} = f C_L V_{DD}^2 . \tag{9.34}$$

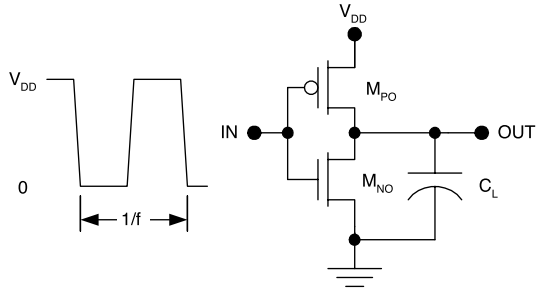


FIGURE 9.9
CMOS gate for consideration of the capacitance switching power.

Normally, the output node of a CMOS gate switches at a frequency lower than the system clock frequency. This is accounted for by introducing an activity factor, α:

$$P_{switch} = \alpha f_{CLK} C_L V_{DD}^2 . \tag{9.35}$$

The capacitance that loads the output node comprises three components:

$$C_L = C_{\text{internal}} + C_{\text{interconnect}} + C_{\text{load}} , \tag{9.36}$$

where C_{internal} is the capacitance internal to the gate associated with the drain junctions of the MOSFETs, $C_{\text{interconnect}}$ is the capacitance of the interconnects, and C_{load} is the input capacitance associated with the load logic gates.

In complex CMOS logic gates internal nodes may switch even when the output node does not. To account for all of the capacitance switching power accurately, it is necessary to sum up the contributions from all of the individual nodes:

$$P_{switch} = f_{CLK} V_{DD} \sum_{i=1}^{N} \alpha_i C_i V_i , \tag{9.37}$$

where α_i is the activity for the i*th* node, C_i is the capacitance loading the i*th* node, and V_i is the average voltage swing on the i*th* node.

9.5.2 Short-Circuit Dissipation

The short-circuit component of the dynamic dissipation arises as a consequence of the simultaneous conduction of the n-MOSFET and p-MOSFET. Therefore, although the capacitance switching power depends only on the voltage swing, the short-circuit dissipation depends also on the rise and fall times at the input and output.

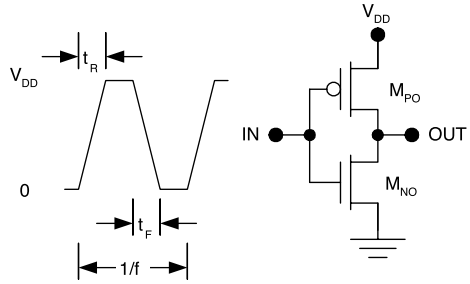


FIGURE 9.10
CMOS inverter for consideration of the short-circuit power.

Consider a symmetric CMOS inverter with a negligible capacitive load as shown in Figure 9.10. For $V_{\text{IN}} \le V_{DD}/2$, the n-MOSFET is saturated and for $V_{\text{IN}} \ge V_{DD}/2$ the p-MOSFET is saturated. The supply current as a function of the input voltage is

$$i_{DD} = \begin{cases} 0; & V_{\text{IN}} \le V_T \\ \dfrac{K}{2}\left(V_{\text{IN}} - V_T\right)^2; & V_{\text{IN}} \le V_{DD}/2 \\ \dfrac{K}{2}\left(V_{DD} - V_{\text{IN}} - V_T\right)^2; & V_{\text{IN}} \ge V_{DD}/2 \\ 0; & V_{\text{IN}} \ge V_{DD} - V_T \end{cases} \tag{9.38}$$

If the rise and fall times for the input waveform are equal:

$$t_R = t_F = \tau\,; \tag{9.39}$$

the time-averaged short-circuit power is

$$\begin{aligned} P_{sc} &= \frac{V_{DD}}{T}\int_0^T i_{DD}\,dt \\ &= \frac{2V_{DD}}{T}\left\{\int_{\tau V_T/V_{DD}}^{\tau/2} \frac{K}{2}\left(\frac{V_{DD}t}{\tau} - V_T\right)^2 dt + \int_{\tau/2}^{\tau-\tau V_T/V_{DD}} \frac{K}{2}\left(V_{DD} - \frac{V_{DD}t}{\tau} - V_T\right)^2 dt\right\} \\ &= \frac{K\tau f\left(V_{DD} - 2V_T\right)^3}{12}. \end{aligned} \tag{9.40}$$

If the switching frequency is less than the clock frequency, it is possible to account for this by invoking the switching activity factor (as in the previous section):

$$P_{sc} = \frac{\alpha K \tau f_{CLK} \left(V_{DD} - 2V_T\right)^3}{12}. \tag{9.41}$$

The short-circuit power increases linearly with the switching frequency and may also be reduced significantly by the reduction of the supply voltage. In fact, it is possible to eliminate the short-circuit power altogether if $V_{DD} < (V_{TN} + |V_{TP}|)$.

9.5.3 Leakage Current Dissipation

The static dissipation in CMOS arises as a consequence of three components of leakage current. These are, in order of importance, the subthreshold current in the MOSFETs, the leakage currents in the source and drain p–n junctions in the MOSFETs, and the leakage in the gate oxide of the MOSFETs.

The subthreshold current in an n-MOSFET with $V_{DS} > 3kT/q$ is given by

$$I_D \approx \frac{\mu_n \varepsilon_{ox} W(1+m)}{t_{ox} L} \left(\frac{kT}{q}\right)^2 \exp\left(\frac{q\left(V_{GS} - V_T\right)}{mkT}\right), \tag{9.42}$$

where

μ_n = electron mobility
ε_{ox} = permittivity of oxide
t_{ox} = oxide thickness
W = width of MOSFET channel
L = length of MOSFET channel
k = Boltzmann constant
T = absolute temperature
q = electronic charge
V_{GS} = gate-to-source bias voltage

The unitless parameter m is given by

$$m = 1 + \frac{C_{dm}}{C_{ox}}, \tag{9.43}$$

where

C_{dm} = maximum depletion layer capacitance of the semiconductor under the oxide
C_{ox} = oxide capacitance

In typical MOSFETs, $1.5 < m < 2$, but for silicon-on-insulator (SOI) MOSFETs, the value of m is approximately unity.

In a CMOS circuit, subthreshold current flows in the n-MOSFET network when the output is high. With a low output, subthreshold current flows in the p-MOSFET network. The average subthreshold current in a symmetric CMOS inverter in either logic state is equal to

$$I_{subthreshold} \approx K(1+m)\left(\frac{kT}{q}\right)^2 \exp\left(-V_T/S\right), \tag{9.44}$$

where the subthreshold swing S is given by

$$S \equiv \left(\frac{d\left(\log_{10} I_D\right)}{dV_{GS}}\right)^{-1} = \frac{mkT}{q} \tag{9.45}$$

and K is the device transconductance parameter of the symmetric n-channel and p-channel MOSFETs. Therefore, the subthreshold power is approximately

$$P_{subthreshold} \approx V_{DD}K(1+m)\left(\frac{kT}{q}\right)^2 \exp\left(-V_T/S\right). \tag{9.46}$$

As a rule of thumb, the absolute value of the threshold voltages should be two to three times the subthreshold swing in order to limit the subthreshold leakage to a tolerable value. In conventional CMOS circuits operating at room temperature, the minimum threshold voltages are therefore 0.3 V. Silicon-on-insulator MOSFETs have near ideal subthreshold characteristics and allow threshold voltages down to 0.1 V.

The reverse-biased p–n junctions in a CMOS circuit also contribute to the leakage dissipation. With a high output from a CMOS circuit, the p-MOSFETs are linear and the reverse-biased drain-body p–n junctions in the n-MOSFETs leak. With a low output, the n-MOSFETs are linear and the drain-body p–n junctions of the p-MOSFETs leak.

The leakage current in a reverse-biased p–n junction is given approximately by a Schockley-type current source,

$$I_{pn} = I_S\left[\exp\left(\frac{qV}{nkT}\right) - 1\right] \approx -I_S. \tag{9.47}$$

Each gate contributes a p–n junction leakage power equal to

$$P_{pn} \approx V_{DD}I_S. \tag{9.48}$$

The gate oxide leakage is due to quantum mechanical tunneling. The theoretical treatment of the oxide leakage current is beyond the scope of this

book. This oxide leakage is negligible for CMOS circuits with gate oxide thicker than about 40 nm. The importance of this contribution is that it places a limit on the scaling of CMOS circuits utilizing silicon dioxide. Other gate insulators with high dielectric constants (high κ dielectrics) are being investigated to extend the scaling limits without undue gate oxide leakage current.

In summary, the dissipation in CMOS circuits is given by

$$P = \underbrace{P_{subthreshold} + P_{pn}}_{P_{DC}} + \underbrace{P_{sc} + P_{switch}}_{P_{AC}}. \tag{9.49}$$

Typically, the dynamic dissipation is dominated by the capacitance switching power and the static dissipation is dominated by the subthreshold conduction. Thus,

$$\begin{aligned} P &\approx P_{switch} + P_{subthreshold} \\ &\approx \alpha f_{CLK} C_L V_{DD}^2 + V_{DD} K (1+m) \left(\frac{kT}{q}\right)^2 \exp\left(-V_T / S\right). \end{aligned} \tag{9.50}$$

Often the subthreshold contribution can be neglected except under conditions of low switching activity, such as standby conditions.

Example 9.4

Calculate and plot the dissipation vs. the switching frequency for a symmetric CMOS gate with $V_{DD} = 2.5$ V, $V_T = 0.6$ V, and $K = 0.5$ mA/V^2 and $C_L = 15$ pF.

Solution. The capacitance switching power is given by $P_{switch} = f(2.5\text{ V})^2 (15 \times 10^{-12}\text{ F}) = f(9.4 \times 10^{-11}\text{ J})$. The short-circuit power can be estimated if the rise and fall times are assumed to be equal to the propagation delay:

$$\tau \approx t_P \approx \frac{15 \times 10^{-12}\text{ F}}{0.5 \times 10^{-3}\text{ A/V}^2} \left[\frac{1.2\text{ V}}{(1.9\text{ V})^2} + \frac{2}{1.9\text{ V}} \ln\left(\frac{1.9\text{ V}}{1.25\text{ V}}\right) \right] = 23\text{ ns},$$

yielding an order-of-magnitude estimate for the short-circuit power:

$$\begin{aligned} P_{sc} &= \frac{K\tau f \left(V_{DD} - 2V_T\right)^3}{12} \\ &= \frac{\left(0.5\text{ mA/V}^2\right)\left(23 \times 10^{-9}\text{ s}\right)\left(2.5\text{ V} - 1.2\text{ V}\right)^3}{12} f, \\ &= \left(2.1 \times 10^{-12}\text{ J}\right) f \end{aligned}$$

which shows that the short circuit power is negligible compared to the capacitance switching power in this case.

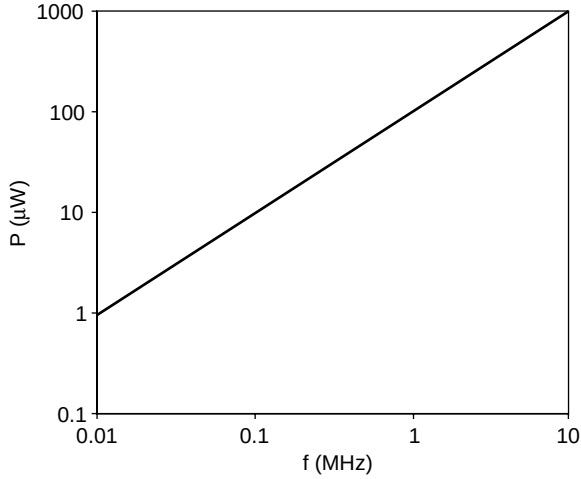


FIGURE 9.11
Example of power vs. switching frequency characteristic for a CMOS gate.

The static dissipation is expected to be dominated by the subthreshold contribution. If it is assumed that m = 1.6, then

$$P_{subthreshold} \approx V_{DD}K(1+m)\left(\frac{kT}{q}\right)^2 \exp\left(-V_T/S\right)$$

$$\approx (2.5\text{ V})\left(0.5\text{ mA/V}^2\right)(1+1.6)(26\text{ mV})^2 \exp\left(-\frac{0.6\text{ V}}{0.1\text{ V}}\right).$$

$$= 5.4\times10^{-9}\text{ W}$$

This contribution is negligible except at very low switching frequencies. Thus, the capacitance switching power dominates and $P \approx P_{switch} = f(9.6 \times 10^{-11}$ J), as plotted in Figure 9.11.

9.6 Fan-Out

The fan-out of CMOS is determined entirely by dynamic considerations. Suppose that a CMOS gate is loaded with similar CMOS circuits. The only DC load current is due to the leakage currents in the MOS gates of the switch transistors, and this is negligible. However, the load gates present a significant capacitive load and thus propagation delays increase with the fan-out. Therefore, the maximum fan-out is dictated by the maximum tolerable propagation delay.

Consider a symmetric CMOS inverter loaded by N similar CMOS gate circuits. The input capacitance for the CMOS load gates is approximately

$$C_{IN} = C_{OXN} + C_{OXP} = \frac{\varepsilon_{OX} W_N L_N}{t_{OX}} + \frac{\varepsilon_{OX} W_P L_P}{t_{OX}}. \tag{9.51}$$

If the maximum allowable propagation delay is $t_{P,\max}$, then the maximum allowable load capacitance is

$$C_{L,\max} \approx \frac{Kt_{P,\max}}{\left[\dfrac{2V_T}{\left(V_{DD} - V_T\right)^2} + \dfrac{2}{\left(V_{DD} - V_T\right)} \ln\left(\dfrac{V_{DD} - V_T}{V_{DD}/2}\right)\right]}. \tag{9.52}$$

The maximum fan-out is the largest integer satisfying

$$N_{MAX} \leq \frac{C_{L,\max}}{C_{IN}}. \tag{9.53}$$

For low-frequency applications the maximum fan-out of a CMOS gate may be in the hundreds or even thousands. For higher-frequency applications, the fan-out should be at least 10. Smaller values greatly restrict the logic design of the system.

9.7 Logic Design

Any desired logic function may be implemented in CMOS by forming parallel and series combinations of MOSFETs. The two-input NAND gate comprises two series n-MOSFETs and two parallel p-MOSFETs as shown in Figure 9.12. The function of this gate is as follows. If 0 V (logic zero) is applied at either input, the associated n-MOSFET turns off while the associated p-MOSFET turns on, bringing the output to V_{DD} (logic one). The output goes low only if both inputs are brought high so that both n-MOSFETs are linear and both p-MOSFETs are cut off. This is the NAND function.

The three-input CMOS NAND gate is realized by placing three n-MOSFETs in series and three p-MOSFETs in parallel, as shown in Figure 9.13. Here too the output will go high if one or more of the inputs go low. The output is brought down to 0 V only if all three inputs are brought high. In general, the M-input CMOS NAND gate requires M n-MOSFETs in series and M p-MOSFETs in parallel, for a total of 2M transistors. In order to maintain the same performance as for the inverter, the n-channel MOSFETs must be scaled

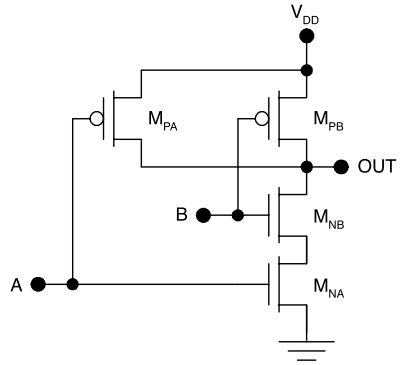


FIGURE 9.12
CMOS NAND2 gate.

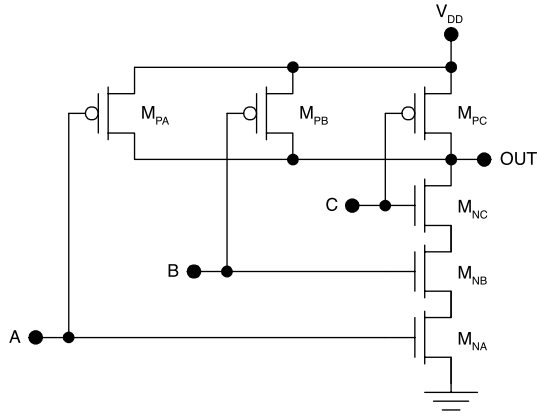


FIGURE 9.13
CMOS NAND3 gate.

by a factor of M (i.e., their *K* values must be increased by a factor of M). This preserves the output low on-state resistance between V_{OUT} and ground, and it preserves the t_{PHL} performance. However, the paralleled p-MOSFETs need not be scaled compared to the inverter.

The two-input CMOS NOR gate is implemented by placing two p-MOSFETs in series and two n-MOSFETs in parallel. The two-input version is shown in Figure 9.14. The function of this circuit is as follows. If logic one (V_{DD}) is applied to either of the inputs, the associated n-MOSFET turns on and the associated p-MOSFET turns off, bringing the output to 0 V. The output goes high only if logic zero is applied to both inputs.

In general, the M-input CMOS NOR gate requires M n-MOSFETs and M p-MOSFETs. The total number of transistors required is 2M, similar to the NAND case. However, *for the NOR gate, the* p-MOSFET*s must be scaled up by a factor of M*. This means that the K values of the p-MOSFETs must be

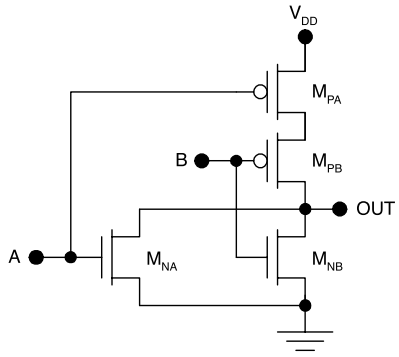


FIGURE 9.14
CMOS NOR2 gate.

increased by a factor of M compared to the inverter with similar performance. In practice, this scaling is achieved by increasing the gate widths while leaving the gate lengths unchanged. The n-MOSFETs need not be scaled in the CMOS NOR gate.

NAND implementations are preferred over NOR implementations in CMOS because of the better area efficiency of the NAND gate. In the M-input NAND gate, the n-MOSFETs are scaled by a factor M whereas in the M-input NOR gate the p-MOSFETs are scaled by a factor of M. Prior to scaling, the p-MOSFETs are already 2.5 times the size of the n-MOSFETs to compensate for the lower mobility of holes compared to electrons. Therefore, for a given level of performance, the M-input NOR gate takes up considerably more chip area than an M-input NAND gate. The system designer can use DeMorgan's theorems to replace NOR gates with NAND gates, thus improving packing density without a penalty in performance.

Complex logic functions can be implemented in CMOS by combining parallel and series branches of n- and p-MOSFETs; an example is shown in Figure 9.15. The logic function performed by this gate is

$$Y = \overline{AB + CD}\,. \tag{9.54}$$

The gate-level representation of this circuit is shown in Figure 9.16.

In CMOS the XOR function is implemented using an AND–OR–INVERT approach. The realization is shown in Figure 9.17 and the gate-level representation appears in Figure 9.18. This circuit performs the function

$$Y = \overline{AB + \overline{A}\,\overline{B}} = A \oplus B\,. \tag{9.55}$$

Implementation of the XOR2 function in CMOS is inefficient, requiring 12 MOSFETs (compared to five MOSFETs in the NMOS realization).

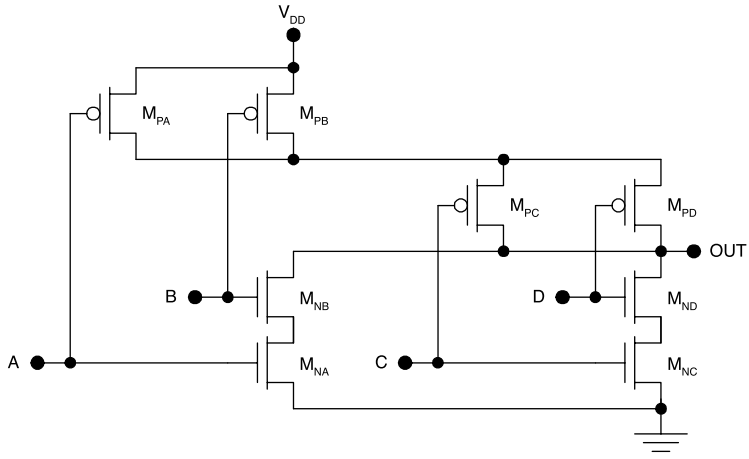


FIGURE 9.15
CMOS AND–OR–INVERT (AOI) gate.

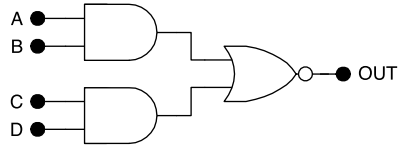


FIGURE 9.16
Gate-level representation of the CMOS logic circuit shown in Figure 9.15.

Other more complex logic functions may be implemented in CMOS by the extension of the NAND and NOR concepts. The individual transistors must be scaled as follows. Any n-MOSFET must be scaled by a factor N, where N is the maximum number of series n-MOSFETs for any path connecting the output and ground, including the n-MOSFET under consideration. Any p-MOSFET must be scaled by a factor M, where M is the maximum number of series n-MOSFETs for any path connecting the output and V_{DD}, and including the p-MOSFET under consideration. The scaling factor may not be the same for all n-MOSFETs and for all p-MOSFETs.

9.8 4000 Series CMOS

An early family of CMOS circuits is the 4000 series.[1] This family of circuits uses aluminum gates, gate lengths of 5 μm, and 100 nm of gate oxide. The

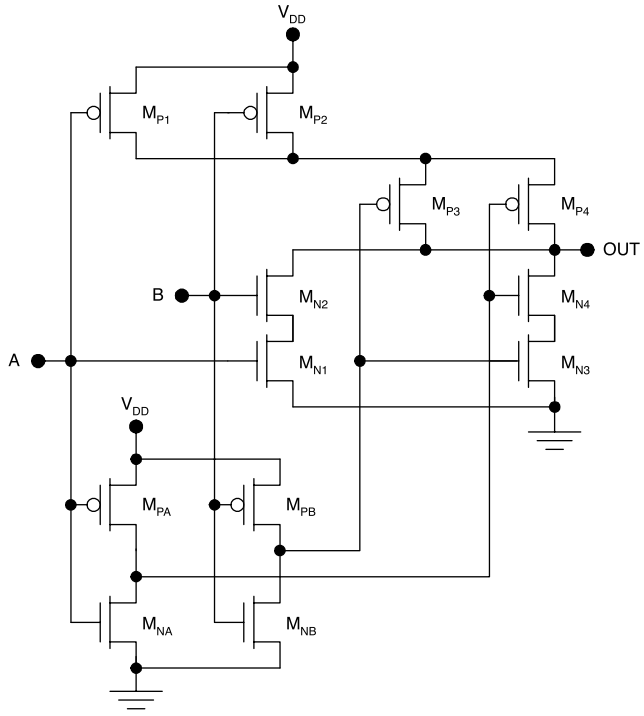


FIGURE 9.17
CMOS XOR2 gate.

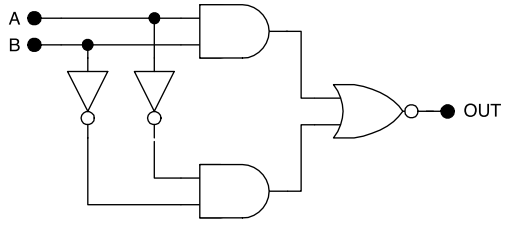


FIGURE 9.18
Gate-level representation of the CMOS XOR2 circuit.

supply voltage can vary in the range from 5 to 15 V. The 4000 series inverter, NAND2, and NOR2 gates are shown in Figure 9.19 through Figure 9.21. The properties of 4000 series CMOS gates are summarized in Table 9.2. The circuits are designed to operate with a wide range of supply voltage from 3 to 15 V. A corresponding range of the propagation delay has a fixed 50-pF load. Because the propagation delay for CMOS varies inversely with the supply voltage, the longest propagation delay of 330 ns corresponds to the lowest supply voltage of 3 V.

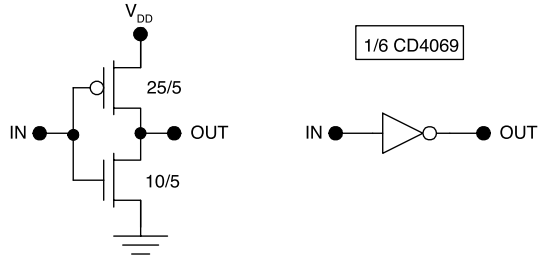


FIGURE 9.19
4000 series CMOS inverter (1/6 of the CD4069 hex inverter). The gate dimensions are given in microns.

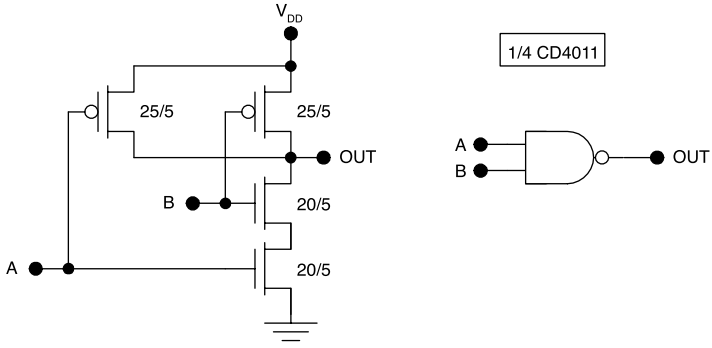


FIGURE 9.20
4000 series CMOS NAND2 gate (1/4 of the CD4011 quad two-input NAND gate). The gate dimensions are given in microns.

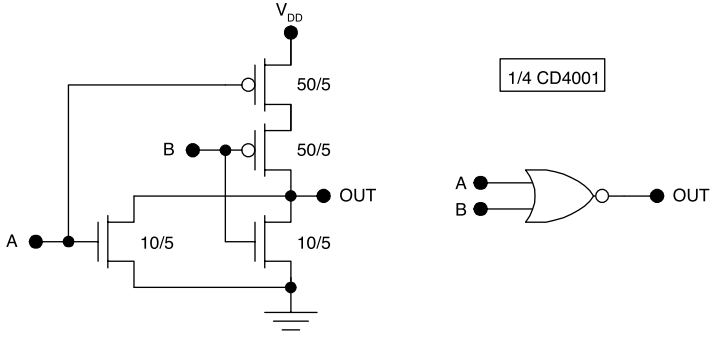


FIGURE 9.21
4000 series CMOS NOR2 gate (1/4 of the CD4001 quad two-input NOR gate). The gate dimensions are given in microns.

TABLE 9.2

Properties of 4000 Series CMOS

4000 series CMOS	
Gate material	Aluminum
Gate length	5 μm
Oxide thickness	100 nm
Supply voltage	3 to 15 V
Propagation delay (C_L = 50 pF)	85 to 330 ns

9.9 74HCxx Series CMOS

The 74HCxx series of CMOS[1–4] was designed to be pin-for-pin compatible with the 74xx series TTL. The circuits were designed with a number of device improvements over 4000 series CMOS, including polysilicon gates, 3-μm gate lengths, and 60-nm thick gate oxide. In addition to these device improvements, the circuit designs were improved by the use of double buffering.

The 74HC00 quad two-input NAND gate comprises four identical circuits like the one shown in Figure 9.22. This circuit is double buffered by two inverters. Although the two inverters do not alter the overall logic function of the circuit, they do provide voltage gain (thus sharpening the voltage transfer characteristic) and current gain (thus allowing the use of smaller input transistors, with reduced input capacitance).

Table 9.3 summarizes the basic characteristics of 74HC high-speed CMOS gates. The range of supply voltages is much tighter than for the 4000 series.

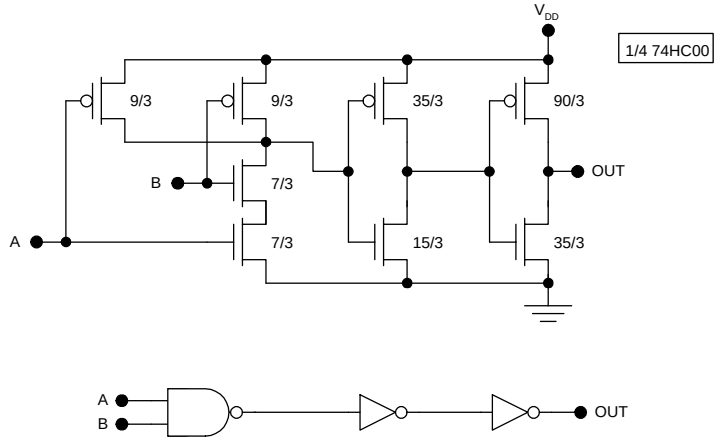


FIGURE 9.22
74HC high-speed CMOS NAND2 gate (1/4 of the 74HC00 quad two-input NAND gate).

TABLE 9.3

Basic Characteristics of High-Speed CMOS[a]

74HC series CMOS	
Gate material	Polysilicon
Gate length	3 μm
Oxide thickness	60 nm
Supply voltage	4.5 to 5.5 V
Propagation delay (C_L = 15 pF)	10 ns

[a] 74HCxx series.

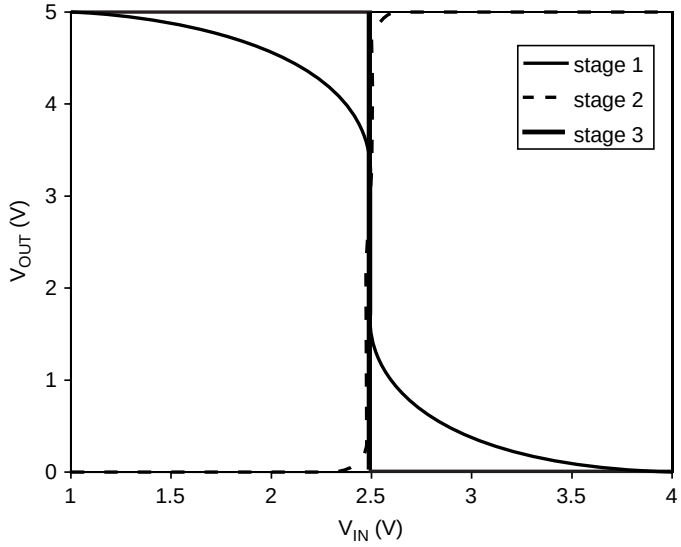


FIGURE 9.23
Voltage transfer characteristics for the first, second, and third (output) stages of the 74HC CMOS inverter (1/6 74HC04).

The upper limit is dictated by the MOSFET breakdown characteristics and the lower limit by the propagation delay specification.

Double buffering improves the voltage transfer characteristics of 74HCxx gates, as shown for the case of the 74HC04 inverter in Figure 9.23. With double buffering, the voltage transfer characteristic becomes ideal, with rail-to-rail voltage swing and an abrupt transition at $V_{DD}/2$. Therefore, $V_{IL} = V_{IH} = V_{DD}/2$ and $V_{NML} = V_{NMH} = V_{DD}/2$. The double buffering does not exact a penalty in terms of the dynamic performance, as will be shown next.

Example 9.5

Estimate the propagation delay for the 74HC CMOS inverter (1/6 74HC04) as shown in Figure 9.24 with a 15-pF load.

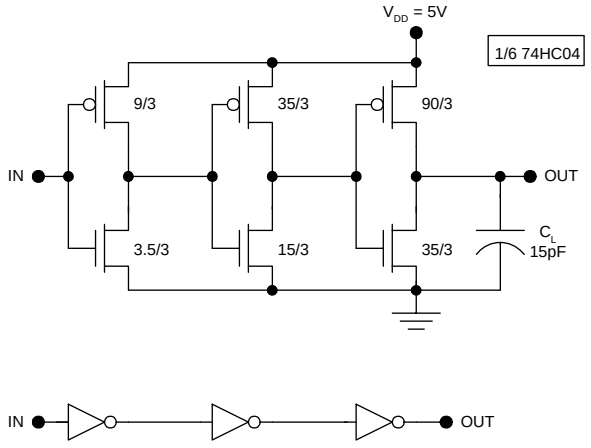


FIGURE 9.24
74HC CMOS inverter (1/6 74HC04) with a 15-pF load.

Solution. The process transconductance parameters for the p-MOSFETs and n-MOSFETs are

$$k'_P = \frac{\mu_p \varepsilon_{ox}}{t_{ox}} = \frac{(230\ \text{cm}^2/\text{Vs})(3.9)(8.85\times10^{-14}\ \text{F/cm})}{60\times10^{-7}\ \text{cm}} = 13.4\ \mu\text{A/V}^2$$

and

$$k'_N = \frac{\mu_n \varepsilon_{ox}}{t_{ox}} = \frac{(550\ \text{cm}^2/\text{Vs})(3.9)(8.85\times10^{-14}\ \text{F/cm})}{60\times10^{-7}\ \text{cm}} = 33.4\ \mu\text{A/V}^2\ ,$$

respectively.

For the first stage, the device transconductance parameters are

$$K_{P1} = k'_P \frac{W_{P1}}{L_{P1}} = 13.4\ \mu\text{A/V}^2\left(\frac{9}{3}\right) = 40\ \mu\text{A/V}^2$$

and

$$K_{N1} = k'_N \frac{W_{N1}}{L_{N1}} = 33.4\ \mu\text{A/V}^2\left(\frac{3.5}{3}\right) = 39\ \mu\text{A/V}^2\ ,$$

respectively. The load capacitance seen by the first stage is the input capacitance for the second stage,

$$C_{L1} = \frac{\varepsilon_{ox} W_{P2} L_{P2}}{t_{ox}} + \frac{\varepsilon_{ox} W_{N2} L_{N2}}{t_{ox}} = 86fF ;$$

the propagation delay for the first stage is therefore

$$t_{P1} \approx \frac{86 \times 10^{-15}\ \text{F}}{40 \times 10^{-6}\ \text{A/V}^2} \left[\frac{1.2\ \text{V}}{(4.4\ \text{V})^2} + \frac{2}{4.4\ \text{V}} \ln\left(\frac{4.4\ \text{V}}{2.5\ \text{V}} \right) \right] = 0.8\ \text{ns} .$$

Similarly, for the second stage, the device transconductance parameters are

$$K_{P2} = k'_P \frac{W_{P2}}{L_{P2}} = 13.4\ \mu\text{A/V}^2 \left(\frac{35}{3} \right) = 0.16\ \text{mA/V}^2$$

and

$$K_{N2} = k'_N \frac{W_{N2}}{L_{N2}} = 33.4\ \mu\text{A/V}^2 \left(\frac{15}{3} \right) = 0.17\ \text{mA/V}^2 .$$

The load capacitance seen by the second stage is the input capacitance for the third stage,

$$C_{L2} = \frac{\varepsilon_{ox} W_{P3} L_{P3}}{t_{ox}} + \frac{\varepsilon_{ox} W_{N3} L_{N3}}{t_{ox}} = 216fF ,$$

and the propagation delay for the second stage is therefore

$$t_{P2} \approx \frac{216 \times 10^{-15}\ \text{F}}{0.16 \times 10^{-3}\ \text{A/V}^2} \left[\frac{1.2\ \text{V}}{(4.4\ \text{V})^2} + \frac{2}{4.4\ \text{V}} \ln\left(\frac{4.4\ \text{V}}{2.5\ \text{V}} \right) \right] = 0.5\ \text{ns} .$$

For the third stage, the device transconductance parameters are

$$K_{P3} = k'_P \frac{W_{P3}}{L_{P3}} = 13.4\ \mu\text{A/V}^2 \left(\frac{90}{3} \right) = 0.4\ \text{mA/V}^2$$

and

$$K_{N3} = k'_N \frac{W_{N2}}{L_{N2}} = 33.4\ \mu\text{A/V}^2 \left(\frac{35}{3} \right) = 0.39\ \text{mA/V}^2 .$$

The load capacitance seen by the third stage is the *external load*,

$$C_{L3} = 15pF\,,$$

so the propagation delay for the third (output) stage is

$$t_{P3} \approx \frac{15\times10^{-12}\ \text{F}}{0.4\times10^{-3}\ \text{A/V}^2}\left[\frac{1.2\ \text{V}}{(4.4\ \text{V})^2}+\frac{2}{4.4\ \text{V}}\ln\left(\frac{4.4\ \text{V}}{2.5\ \text{V}}\right)\right]=13.8\ \text{ns}\,.$$

The overall propagation delay for the 74HC04 inverter with a 15-pF external load can be found by adding the individual propagation delays of the three stages:

$$t_P = t_{P1}+t_{P2}+t_{P3} = 0.8\ \text{ns}+0.5\ \text{ns}+13.8\ \text{ns} = 15.1\ \text{ns}\,.$$

Therefore, the output stage accounts for 90% of the overall propagation delay.

9.10 Buffered CMOS

A modern VLSI CMOS chip contains a million or more gates but only ~10^3 external connections. Therefore most of the gates experience only on-chip loads and are capable of propagation delays in picoseconds without the need for buffering. For the externally connected gates, significant capacitive loading exists, so buffering is necessary to reduce the propagation delays and achieve acceptable off-chip data rates.

Example 9.6

Estimate the propagation delay for a single CMOS stage driving a 50-pF off-chip load as shown in Figure 9.25. Assume quarter micron CMOS technology using a 1.8-V supply for all gates, including the output drivers, and ±0.5-V threshold voltages. For this technology, the inverters internal to the chip have gate dimensions of 2.2 μm/0.25 μm (p-MOSFET) and 0.9 μm/0.25 μm (n-MOSFET) and the oxide thickness is 7 nm.

Solution. The process transconductance parameters for this technology are

$$k_P' = \frac{\mu_p\varepsilon_{ox}}{t_{ox}} = \frac{(230\ \text{cm}^2/\text{Vs})(3.9)(8.85\times10^{-14}\ \text{F/cm})}{7\times10^{-7}\ \text{cm}} = 0.113\ \text{mA/V}^2$$

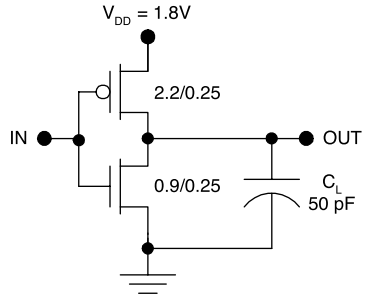


FIGURE 9.25
A single stage of CMOS driving a 50-pF off-chip load.

and

$$k'_N = \frac{\mu_n \varepsilon_{ox}}{t_{ox}} = \frac{(580\ \text{cm}^2/\text{Vs})(3.9)(8.85\times10^{-14}\ \text{F/cm})}{7\times10^{-7}\ \text{cm}} = 0.28\ \text{mA/V}^2\ ,$$

respectively.

Now suppose that a single (unbuffered) stage drives an external 50-pF load as shown in Figure 9.25. For the single stage, the device transconductance parameters are

$$K_P = k'_P \frac{W_P}{L_P} = 0.113\ \text{mA/V}^2 \left(\frac{2.2}{0.25}\right) = 0.99\ \text{mA/V}^2$$

and

$$K_N = k'_N \frac{W_N}{L_N} = 0.28\ \text{mA/V}^2 \left(\frac{0.9}{0.25}\right) = 1.01\ \text{mA/V}^2\ ,$$

respectively. The propagation delay for the single stage is

$$t_{P1} \approx \frac{50\times10^{-12}\ \text{F}}{0.97\ \text{mA/V}^2}\left[\frac{1.0\ \text{V}}{(1.3\ \text{V})^2} + \frac{2}{1.3\ \text{V}}\ln\left(\frac{1.3\ \text{V}}{0.9\ \text{V}}\right)\right] = 58\ \text{ns}\ .$$

Example 9.7

Estimate the propagation delay for three stages of CMOS driving a 50-pF off-chip load, with a scaling factor of five applied at each stage of buffering, as depicted in Figure 9.26. Assume quarter micron CMOS technology using a 1.8-V supply for all gates, including the output drivers, and ±0.5-V threshold voltages. For this technology, the inverters internal to the chip have gates

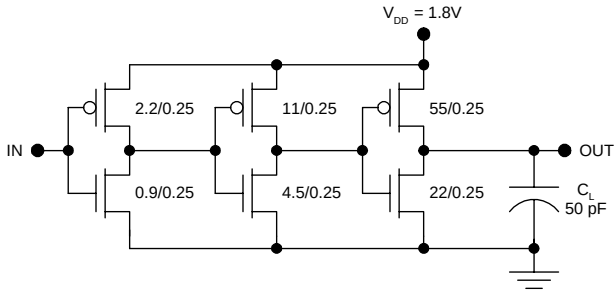


FIGURE 9.26
Three stages of CMOS driving a 50-pF off-chip load.

dimensions of 2.2 μm/0.25 μm (p-MOSFET) and 0.9 μm/0.25 μm (n-MOSFET) and the oxide thickness is 7 nm.

Solution. From one stage to the next, both gate widths have been scaled up by a factor of five. For the first stage, the device transconductance parameters are

$$K_{P1} = k'_P \frac{W_{P1}}{L_{P1}} = 0.113\ \text{mA/V}^2 \left(\frac{2.2}{0.25}\right) = 0.99\ \text{mA/V}^2$$

and

$$K_{N1} = k'_N \frac{W_{N1}}{L_{N1}} = 0.28\ \text{mA/V}^2 \left(\frac{0.9}{0.25}\right) = 1.01\ \text{mA/V}^2\ .$$

The load capacitance for the first stage is equal to the input capacitance of the second stage,

$$C_{L1} = \frac{\varepsilon_{OX} W_{P2} L_{P2}}{t_{OX}} + \frac{\varepsilon_{OX} W_{N2} L_{N2}}{t_{OX}} = 19.1fF\ ;$$

therefore, the propagation delay for the first stage is

$$t_{P1} \approx \frac{19.1\times10^{-15}\,\text{F}}{0.99\ \text{mA/V}^2}\left[\frac{1.0\ \text{V}}{(1.3\ \text{V})^2} + \frac{2}{1.3\ \text{V}}\ln\left(\frac{1.3\ \text{V}}{0.9\ \text{V}}\right)\right] = 23\ \text{ps}\ .$$

For the second stage,

$$K_{P2} = k'_P \frac{W_{P2}}{L_{P2}} = 0.113\ \text{mA/V}^2 \left(\frac{11}{0.25}\right) = 5.0\ \text{mA/V}^2\ ,$$

$$K_{N2} = k'_N \frac{W_{N2}}{L_{N2}} = 0.28\ \text{mA/V}^2 \left(\frac{4.5}{0.25}\right) = 5.0\ \text{mA/V}^2,$$

$$C_{L2} = \frac{\varepsilon_{OX} W_{P3} L_{P3}}{t_{OX}} + \frac{\varepsilon_{OX} W_{N3} L_{N3}}{t_{OX}} = 96fF,$$

and

$$t_{P2} \approx \frac{96 \times 10^{-15}\ \text{F}}{5.0\ \text{mA/V}^2} \left[\frac{1.0\ \text{V}}{(1.3\ \text{V})^2} + \frac{2}{1.3\ \text{V}} \ln\left(\frac{1.3\ \text{V}}{0.9\ \text{V}}\right)\right] = 23\ \text{ps}.$$

For the third stage,

$$K_{P3} = k'_P \frac{W_{P3}}{L_{P3}} = 0.113\ \text{mA/V}^2 \left(\frac{55}{0.25}\right) = 25\ \text{mA/V}^2,$$

$$K_{N3} = k'_N \frac{W_{N3}}{L_{N3}} = 0.28\ \text{mA/V}^2 \left(\frac{22}{0.25}\right) = 25\ \text{mA/V}^2,$$

$$C_{L3} = C_L = 50\ \text{pF},$$

and

$$t_{P3} \approx \frac{50 \times 10^{-12}\ \text{F}}{25\ \text{mA/V}^2} \left[\frac{1.0\ \text{V}}{(1.3\ \text{V})^2} + \frac{2}{1.3\ \text{V}} \ln\left(\frac{1.3\ \text{V}}{0.9\ \text{V}}\right)\right] = 2.3\ \text{ns}.$$

The total propagation delay for the case of three stages is $t_P = t_{P1} + t_{P2} + t_{P3}$ = 23 ps + 23 ps + 2.3 ns = 2.34 ns.

Therefore, the use of two buffer stages reduced the propagation delay by a factor of approximately 1/25. This is because the buffer stages increased the current-driving capability by 25 without adding appreciable propagation delays of their own.

Example 9.8

Estimate the propagation delay for five stages of CMOS driving a 50-pF off-chip load as shown in Figure 9.27, with a scaling factor of five applied at each stage of buffering. Assume quarter micron CMOS technology using a 1.8-V supply for all gates, including the output drivers, and ±0.5-V threshold voltages. For this technology, the inverters internal to the chip have gate dimensions of 2.2 μm/0.25 μm (p-MOSFET) and 0.9 μm/0.25 μm (n-MOSFET) and the oxide thickness is 7 nm.

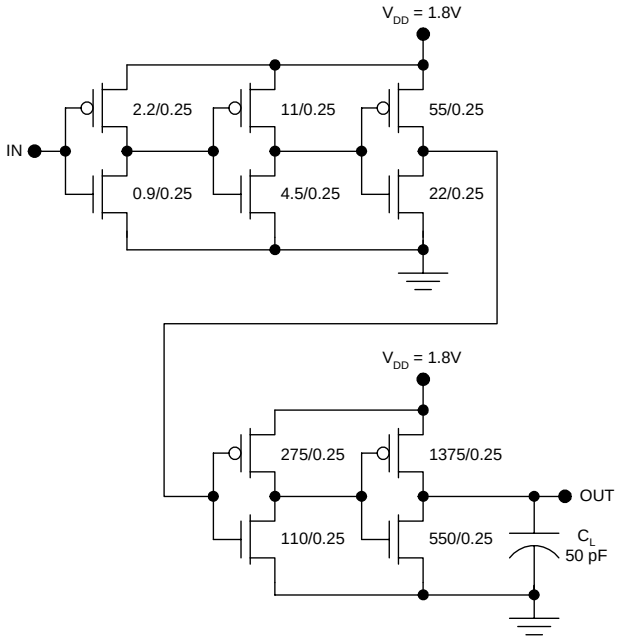


FIGURE 9.27
Five stages of CMOS driving a 50-pF off-chip load.

Solution. Here, the first stage has the same device transconductance parameters and load as those in the previous two examples; therefore, its propagation delay is unchanged at 23 ps. For stages two through four, the individual propagation delays are each 23 ps. This is because each successive stage experiences a fivefold increase in the load capacitance, which is compensated for by a fivefold increase in the device transconductance parameters. For the fifth and final stage,

$$K_{P5} = k'_P \frac{W_{P5}}{L_{P5}} = 0.113 \text{ mA/V}^2 \left(\frac{1375}{0.25}\right) = 620 \text{ mA/V}^2 ,$$

$$K_{N5} = k'_N \frac{W_{N5}}{L_{N5}} = 0.28 \text{ mA/V}^2 \left(\frac{550}{0.25}\right) = 620 \text{ mA/V}^2 ,$$

$$C_{L5} = C_L = 50 \text{ pF} ,$$

and

$$t_{P5} \approx \frac{50 \times 10^{-12} \text{ F}}{620 \text{ mA/V}^2} \left[\frac{1.0 \text{ V}}{(1.3 \text{ V})^2} + \frac{2}{1.3 \text{ V}} \ln\left(\frac{1.3 \text{ V}}{0.9 \text{ V}}\right) \right] = 93 \text{ ps} .$$

The total propagation delay for the case of five stages (four buffers) is

$$t_P = \sum_{n=1}^{5} t_{Pn} = 4(23\text{ ps}) + 93\text{ ps} = 185\text{ ps}\,.$$

Here, the total propagation delay has been reduced almost by a factor of $1/5^5$ with five buffer stages.

Based on the previous examples, it might be tempting to try to reduce the scaling factor and achieve shorter propagation delays at each stage. However, if the scaling factor is too small, the added delay of the stage more than compensates for the added current-driving capability. It has been shown that the optimum scaling factor is therefore "e." In common practice, this factor is rounded off to 3.

It is important to note that buffering involves a trade-off between performance and chip area. In the five-stage case just considered, the p-MOSFET of the driving stage is so large that it could be seen by the naked eye. Clearly, this buffer scheme will take up a tremendous amount of chip area, which then cannot be used for adding chip functionality. Nonetheless, it has become common practice to use multiple buffering stages in modern CMOS gates. The need for high off-chip data rates outweighs consideration of chip area.

9.11 Pseudo NMOS

In CMOS technology it is possible to implement NMOS-type circuitry using a p-channel MOSFET load. This type of logic circuitry is called "pseudo NMOS" and it has three main advantages. First, like NMOS, pseudo NMOS achieves *higher packing density* when compared to CMOS. Second, pseudo NMOS only requires the fabrication of enhancement type n-channel MOSFETs, so it is *compatible with CMOS fabrication technology.* Third, pseudo NMOS allows *hard-wiring outputs* together to form new logic functions. This approach, called wired logic, is described in Chapter 13 and has potential benefits in packing density and power dissipation.

A pseudo NMOS inverter is depicted in Figure 9.28. The circuit is identical to an NMOS inverter with the exception of the load, which is an enhancement type p-channel MOSFET. Therefore, this circuit is compatible with CMOS fabrication technology, which does not provide for depletion type n-channel MOSFETS. The gate of the p-channel load is connected to ground rather than the input, as in CMOS. Therefore, the pull-up is effectively *passive* so that wired logic can be exploited as described in Chapter 13.

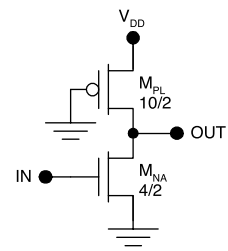


FIGURE 9.28
Pseudo NMOS inverter.

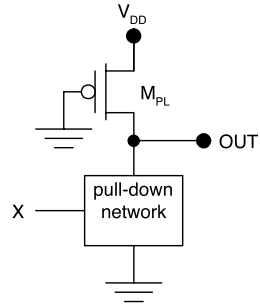


FIGURE 9.29
General pseudo NMOS logic gate.

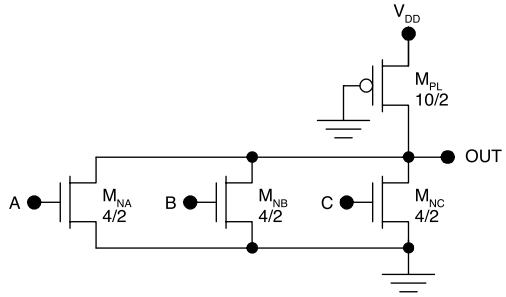


FIGURE 9.30
Pseudo NMOS NOR3 gate.

Pseudo NMOS allows realization of arbitrary logic functions by addition of transistors to the pull-down network, as shown in Figure 9.29. Here, X is a general input vector comprising any number of scalar inputs; the pseudo NMOS concept allows elimination of the p-MOSFET pull-up network. The pull-up network is replaced by a single p-channel MOSFET, regardless of the complexity of the logic function. For the pull-down network, the circuit design and scaling principles are identical to the case of NMOS. As an example, consider the three-input NOR (NOR3) gate shown in Figure 9.30. Each electrical path from the output to ground includes just one n-channel MOSFET; thus, these transistors need not be scaled compared to the inverter. The p-channel load transistor does not require scaling either. This contrasts with the CMOS NOR3 circuit, in which three series p-channel transistors must each be scaled up in width by a factor of three. Clearly the pseudo NMOS circuitry will take less chip area and achieve higher packing densities compared to CMOS.

Implementation of the XOR2 function is rather efficient in pseudo NMOS, as shown in Figure 9.31. Generally, the realization of a logic function involving N inputs requires N + 1 transistors in pseudo NMOS but 2N transistors in CMOS. When the scaling of transistors is accounted for, the packing density of pseudo NMOS can exceed that of CMOS by a factor of four, a

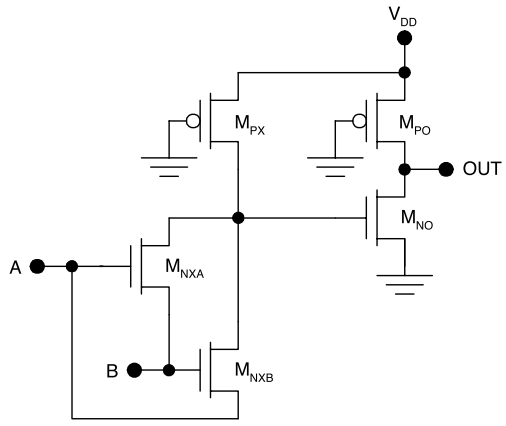


FIGURE 9.31
Pseudo NMOS XOR2 gate.

significant advantage. The disadvantage of pseudo NMOS is the static power dissipation: under output low conditions, a steady DC current will flow in the p-channel load and pull-down network.

9.12 Dynamic CMOS

Dynamic, or clocked, CMOS gates achieve higher packing densities and lower dissipation than the static CMOS gates discussed so far.[5–7] The principle underlying the operation of dynamic CMOS is that the CMOS circuit nodes exhibit capacitance and can be charged to a particular voltage level. Once charged, such a node can be read at a later time as long as the stored charge has not leaked away.

This principle is used in the dynamic CMOS inverter shown in Figure 9.32. When the clock signal goes low, the precharge transistor, M_{PPRE}, turns on and charges the output node to V_{DD}. Then, when the clock signal goes high, the evaluate transistor turns on and allows the output node to discharge if and only if the transistor M_{NO} is also on. Thus, the output signal evaluated at the end of the evaluate interval will be the inversion of the input signal. This is demonstrated in the timing diagram of Figure 9.33.

Notice that, in Figure 9.33, the output voltage dips slightly when the input waveform makes the low-to-high transition. This phenomenon is a consequence of charge sharing and can be understood with the benefit of Figure 9.34. Suppose the output capacitor C_Y is initially precharged to a voltage of V_{DD}, but the evaluation capacitance has zero voltage on it. Then, if the input makes a low-to-high transition, turning on M_{NO}, the electrical charge initially

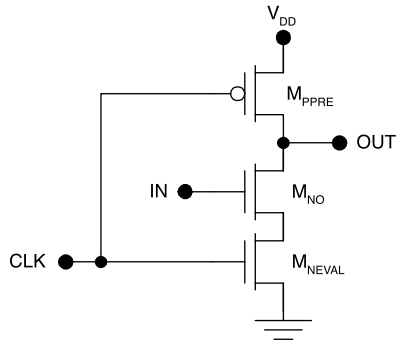


FIGURE 9.32
Dynamic CMOS inverter.

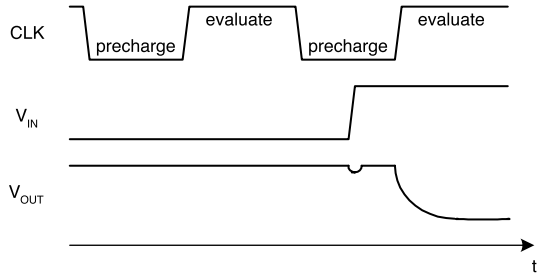


FIGURE 9.33
Waveforms for the dynamic CMOS inverter.

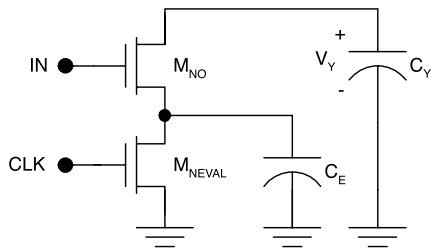


FIGURE 9.34
Charge sharing in a dynamic CMOS inverter (precharge transistor not shown).

present on C_Y will be shared between the two capacitances. The final value of V_Y will be

$$V_Y = \frac{V_{DD} C_Y}{C_E + C_Y} . \tag{9.56}$$

In a static CMOS circuit, charge sharing is of no consequence because the output is actively pulled up or down at every point in time. In contrast, charge sharing is a critical design consideration with dynamic CMOS.

The output voltage of a dynamic CMOS deteriorates very slowly with time because of the low leakage currents associated with CMOS circuitry, but eventually the signal must be refreshed. This necessitates that the clock run at some minimum frequency in order to maintain signal integrity — a disadvantage compared to static CMOS, which can be slowed down arbitrarily or even stopped. Yet another disadvantage of dynamic CMOS is the need for the clock in the first place. In VLSI circuits, the clock presents the most difficult routing design. This has led to active research in the field of free-running asynchronous chips that avoid the clock.

Other logic functions can be realized in dynamic CMOS using a single precharge transistor and a single evaluate transistor. The general structure of a dynamic CMOS gate with a pull-down network is shown in Figure 9.35. Here X is an input vector comprising N scalar inputs. The circuit design and scaling rules are the same as for the design of the pull-down network in an NMOS gate. Therefore, with N inputs the dynamic CMOS gate requires N + 2 transistors, compared to 2N transistors for the static CMOS gate. This difference, along with the scaling requirements imposed on the p-MOSFETs in static CMOS, accounts for the higher packing densities in dynamic CMOS logic.

Figure 9.36 shows a three-input dynamic CMOS NOR gate and Figure 9.37 shows a two-input dynamic CMOS NAND gate. The precharge and evaluate transistors are designed with the same dimensions in either case. This is true for more complicated logic functions as well. In the NOR3 gate, the transistors in the pull-down network act in parallel so they need not be scaled up from the case of the inverter. In the NAND2 gate, the electrical pull-down path from the output to the evaluate transistor involves two series n-MOSFETs. Thus, each of these transistors must be scaled up by a factor of two compared

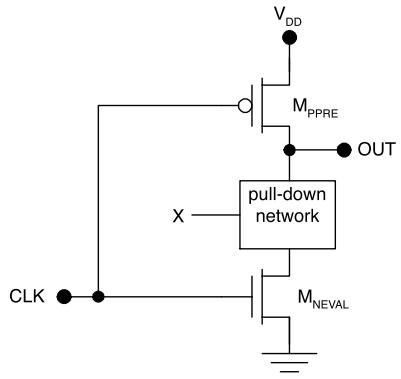


FIGURE 9.35
Dynamic CMOS gate with a pull-down network.

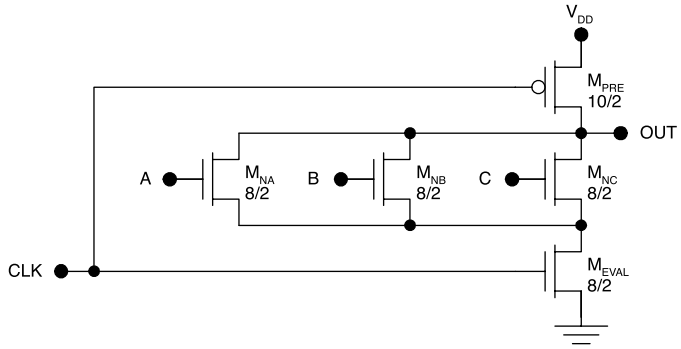


FIGURE 9.36
Dynamic CMOS NOR3 gate.

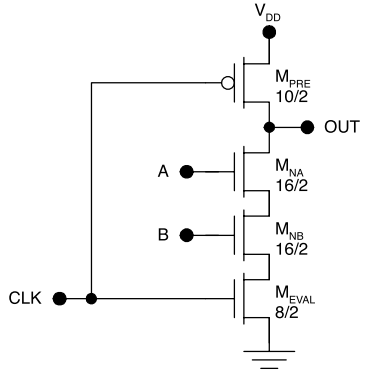


FIGURE 9.37
Dynamic CMOS NAND2 gate.

to the inverter. More complicated logic functions may be realized also, as long as the transistors in the pull-down network are scaled appropriately.

An alternative form of dynamic CMOS uses a p-MOSFET pull-up network as shown in Figure 9.38. Therefore, the pull-up or the pull-down network may be eliminated, with a saving in chip area. Either realization requires N + 2 transistors for the case of N inputs. However, the pull-down realization is more efficient and usually preferred because of the smaller width of n-MOSFETs compared with p-MOSFETs for the same transconductance.

The most serious limitation of dynamic CMOS in VLSI applications is that these circuits *cannot be cascaded*. For the sake of illustration, consider two cascaded dynamic CMOS inverters as shown in Figure 9.39. The behavior of this circuit can be understood using the waveforms of Figure 9.40. Referring to Figure 9.40, the B and OUT nodes are precharged to V_{DD} during the first precharge interval of the clock (clock low). During the first evaluate interval (clock high), M_{NOA} is cut off, so the B node remains at V_{DD} but the

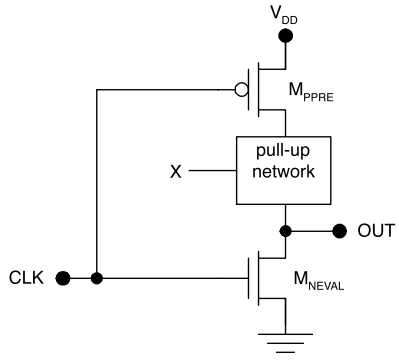


FIGURE 9.38
Dynamic CMOS with a pull-up network.

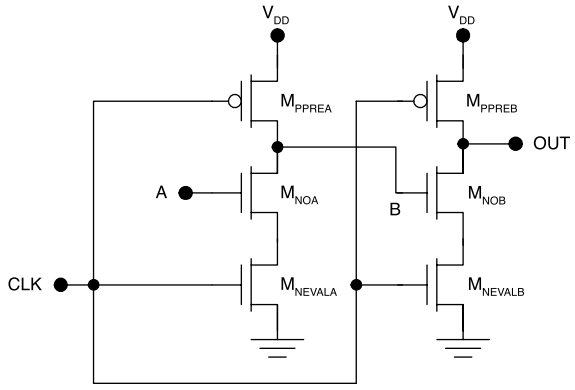


FIGURE 9.39
Cascaded dynamic CMOS inverters.

OUT node discharges when M_{NOB} turns on. Therefore, during the first evaluate interval, OUT = A as required.

During the next precharge interval, B remains at V_{DD} and OUT is charged up to V_{DD}. Suppose that midway through this precharge interval, the A input makes a transition from low to high. This causes a dip in the voltage at the B node due to charge sharing. During the next evaluate interval, M_{NOA} turns on and the B node discharges to zero; therefore, the first cascaded inverter functions as expected. However, node B does not discharge instantaneously. During the time that it takes for node B to discharge, M_{NOB} will conduct so that the output voltage will drop significantly below V_{DD}. This results in a spurious output signal that may not be interpreted as logic "one" by another gate. Therefore, in dynamic circuits it is necessary that a circuit finish its evaluation phase before the circuit in the next stage starts its evaluation phase. This can be achieved using a type of dynamic circuit referred to as *domino logic*.

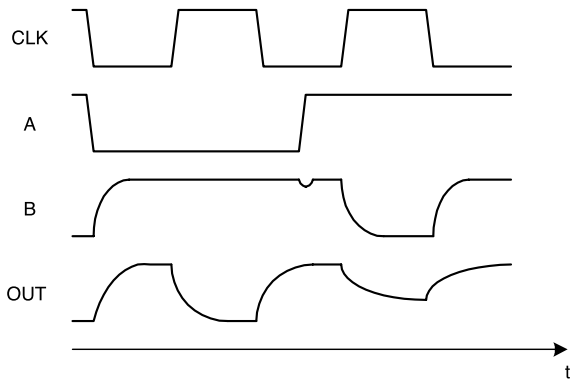


FIGURE 9.40
Waveforms for the cascaded dynamic CMOS inverters.

9.13 Domino Logic

In domino logic,[6,7] the circuits are designed so that a gate does not begin to evaluate until the gates in the previous stage have finished evaluating. This is achieved by two modifications to the dynamic CMOS described in the previous section:

- Gates alternate between the use of pull-down and pull-up networks.
- Gates with pull-up networks use an inverted clock signal.

Consider the three stages of domino logic shown in Figure 9.41. The first stage uses an n-MOSFET pull-down network, the second uses a p-MOSFET pull-up network and an inverted clock, and the third stage uses an n-MOSFET pull-down network. More stages can be added but a pull-down stage always drives a pull-up stage and a pull-up stage always drives a pull-down stage. All pull-up stages use the inverted clock signal and in these stages the *n-MOSFET* performs the *precharge* function, whereas the p-MOSFET performs the *evaluation* function.

When the CLK signal goes low, the outputs of the first and third stages precharge to V_{DD}, while the output of the second stage precharges to zero. Because the output of the first stage is precharged to V_{DD}, the p-MOSFETs in the second stage are held off during the precharging phase for the second stage. Similarly, with the output of the second stage precharged to zero, the n-MOSFETs in the third stage will be held off during the precharging phase. In this way, signal integrity is maintained from one stage to the next and an arbitrary number of stages may be cascaded in this way.

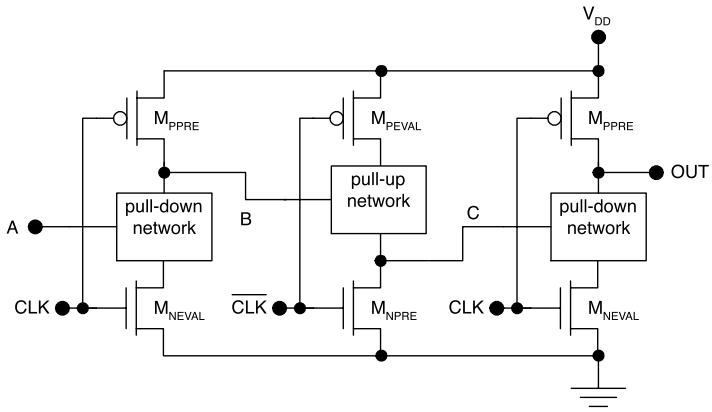


FIGURE 9.41
Three stages of domino logic.

Other realizations of domino logic are possible as long as orderly evaluation is insured. Therefore, signals propagate through the system one stage per clock cycle; the clock period must be long enough to accommodate the longest propagation delay in the system. To some extent, therefore, the clock limits the speed of domino logic.

9.14 Latch-Up in CMOS

In integrated form, complementary pairs of n-MOSFETs and p-MOSFETs contain parasitic bipolar junction transistors. Worse yet, these bipolar transistors appear in npn–pnp pairs, which form two pnpn thyristors. Once turned on, either of these thyristors will latch on, short out the power supply, and maintain an excessive flow of current until the CMOS circuit is destroyed. This problem, called "latch-up," may be prevented by careful design of the CMOS transistors.[8–11]

The four parasitic bipolar transistors of an integrated CMOS inverter are shown in Figure 9.42, which has two parasitic npn transistors and two parasitic pnp transistors. Although the illustration shows a dual-well CMOS structure, other designs are qualitatively similar. The latch-up problem can be better understood by the construction of the equivalent circuit for the CMOS inverter, including the parasitics (Figure 9.43). Notice that Q_2 and Q_4 form one thyristor and Q_1 and Q_3 form a second.

Two disastrous latch-up conditions are possible in this circuit. First, with logic-zero input, it is possible for Q_1 and Q_3 to latch on while M_{PO} is operating in the linear mode. This shorts the power supply through these three devices, allowing destructive levels of current to flow. Second, with logic-one input,

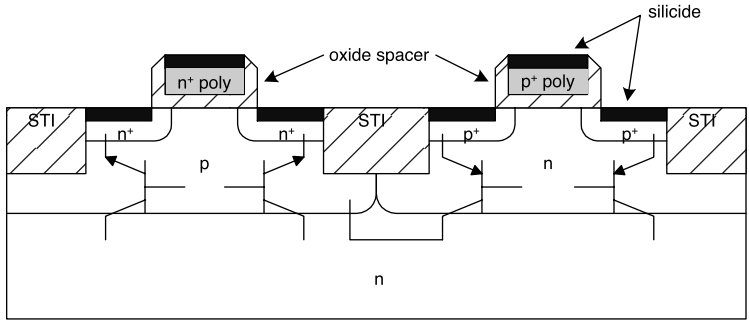


FIGURE 9.42
Parasitic bipolar junction transistors in an integrated CMOS structure using a dual-well structure (STO = shallow trench oxide).

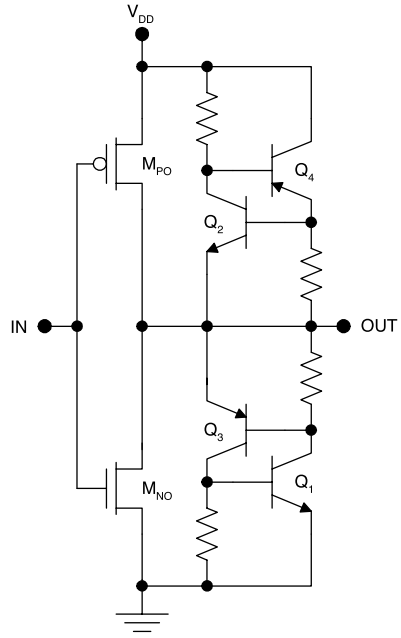


FIGURE 9.43
CMOS inverter equivalent circuit with parasitic bipolar transistors.

it is possible for Q_2 and Q_4 to latch on while M_{NO} is conducting. This shorts the power supply through these three devices, also resulting in destructive levels of current. Latch-up may be prevented by reducing the current gains in the parasitic transistors, which is achieved by adjusting the dimensions or doping in their bases. In modern CMOS designs, the use of epitaxial layers and shallow trench isolation has practically eliminated latch-up.

9.15 Static Discharge in CMOS

MOSFETs are susceptible to damage by static discharge; this is especially true for the scaled-down devices used in modern CMOS circuits. Static discharge involves the destructive breakdown of the gate oxide and must be avoided. For gate-quality silicon dioxide, this destructive breakdown occurs with an electric field intensity of about 10 million V per centimeter. Although seemingly large, such an electric field can result from casual contact with people. Static electricity buildup on the people handling CMOS chips is a natural consequence of ordinary contact with furniture, carpets, clothing, hair, or other materials and is especially problematic in dry environments such as air-conditioned buildings.

Example 9.9

For a 2 μm × 0.25 μm gate with 5 nm oxide, calculate the static charge that will result in destructive breakdown.

Solution. The oxide will break down with an applied voltage of V = (10^7 V/cm) (5×10^{-7} cm) = 5 V. This occurs with the application of a static charge equal to

$$Q = \left(\frac{(2\times10^{-4}\ \text{cm})(0.25\times10^{-4}\ \text{cm})(3.9)(8.85\times10^{-14}\ \text{F/cm})}{5\times10^{-7}\ \text{cm}} \right) 5\ \text{V} = 1.66\times10^{-14}\ \text{C},$$

corresponding to only 100,000 electrons; a person can readily develop a static charge of 10 times this amount.

A number of precautions can be taken to protect CMOS parts from static discharge. First, workers handling the chips should wear ground straps* to bleed off any static charge they may develop. During shipping, CMOS circuits are shipped in conductive packaging, which shorts all input leads to the ground pin. Circuit board assembly requires the use of grounding straps for all workers; however, once the chips are soldered into circuit board assemblies they are relatively safe.

Additional protection is also built into the chips[12–18]; one such scheme is shown in Figure 9.44. The bonding pad is a large metal pad used for bonding gold wires that make the connections to the outside. In flip-chip packages, the bonding pad is replaced by a solder bump but the underlying approach to static protection remains unchanged. Essentially, p–n junction devices are used to limit the voltage applied to any CMOS gate. In the scheme shown, the diode prevents the CMOS input voltage from swinging more than 0.7 V above V_{DD}. The diode-connected bipolar transistor prevents the CMOS input voltage from swinging more negative than about –0.7 V. Excess voltage

* These ground straps are normally worn on wrists and include 1 MΩ resistors for safety.

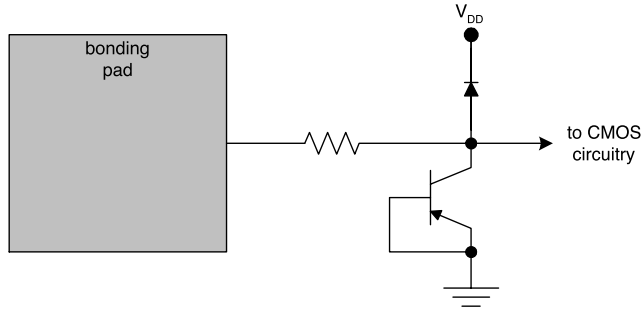


FIGURE 9.44
Static protection circuit for CMOS.

applied at the bonding pad is dropped in the series resistance ever present in the metal interconnect.

9.16 Scaling of CMOS

Over the past three decades scaling of CMOS devices has yielded tremendous gains in chip performance and functionality.[19–28] Scaling refers to the systematic reduction of transistor dimensions from one generation to the next, reducing the parasitic capacitances and also the carrier transit times in the devices. One result of this scaling is an improvement in circuit speed; this improvement is quite dramatic in the case of MOSFET-based circuits. Therefore, another result of scaling is that it narrows the performance gap between CMOS and logic gates based on bipolar transistors. Finally, reduction of the transistor dimensions improves the packing density of CMOS. Remarkably, engineers have doubled the number of transistors per chip every 18 months for the last four decades!

9.16.1 Full Scaling of CMOS

There are many possible schemes for scaling CMOS. Of these, the most obvious approach is "full scaling," which involves scaling all dimensions and voltages by the same factor, 1/s, where s is greater than one. For example, if a scaling factor of $1/\sqrt{2}$ is used ($s = \sqrt{2}$), then the packing density in transistors per square centimeter will be doubled. The motivation for scaling the voltages is that this will leave the electric field intensities unchanged, thus avoiding breakdown effects. Table 9.4 provides a description of full scaling of CMOS. The key dimensions that scale are gate dimensions and oxide thickness; the key voltage that scales is the supply voltage. This table

TABLE 9.4

Full Scaling of CMOS

Parameter	Relationship	Scales by
L	—	$1/s$
W	—	$1/s$
t_{ox}	—	$1/s$
V_{DD}	—	$1/s$
C_{OX}	$\frac{\varepsilon_{ox} WL}{t_{ox}}$	$1/s$
K	$\frac{W}{L}\frac{\mu\varepsilon_{ox}}{t_{ox}}$	s
t_P	$\propto \frac{C_L}{V_{DD}K} \propto \frac{C_{OX}}{V_{DD}K}$	$1/s$
P	$fC_L V^2_{DD}$	$1/s^3$
Packing density	—	s^2
Power density	—	$1/s$

Note: Analysis is based on the assumptions of on-chip loads with constant fan-out and constant clock frequency.

also provides an analysis of how full scaling affects performance that is based on the assumption of similar on-chip loads with a constant fan-out (the load capacitance is assumed to scale with the input capacitance). The benefits of full scaling are improved circuit speed (t_P scales by $1/s$) and packing density (the packing density scales by s^2).

9.16.2 Constant Voltage Scaling of CMOS

Another possible approach to scaling of CMOS is called "constant voltage scaling." This involves scaling all dimensions by the factor $1/s$ while all voltages are kept constant. Although constant voltage scaling increases the electric field intensities in the devices, it is desirable simply because it allows the customer to keep the same supply voltage from one generation to the next. Table 9.5 provides a description and analysis of constant voltage scaling. As in the preceding section, the analysis is based on the assumption of similar on-chip loads with a constant fan-out.

In terms of packing density, constant voltage scaling provides the same benefit as full scaling. However, when it comes to performance, *constant voltage scaling is better than full scaling in terms of the benefit of speed performance* (t_P scales by $1/s^2$). On the other hand, constant voltage scaling is accompanied by a large and undesirable jump in the power density. (Power density scales as s.) Indeed, repeated application of constant voltage scaling over many generations has brought CMOS circuits to the point at which microprocessors

TABLE 9.5

Constant Voltage Scaling of CMOS

Parameter	Relationship	Scales by
L	—	1/s
W	—	1/s
t_{ox}	—	1/s
V_{DD}	—	Unchanged
C_{OX}	$\frac{\varepsilon_{ox} WL}{t_{ox}}$	1/s
K	$\frac{W}{L}\frac{\mu\varepsilon_{ox}}{t_{ox}}$	s
t_P	$\propto \frac{C_L}{K} \propto \frac{C_{OX}}{K}$	$1/s^2$
P	$fC_L V^2_{DD}$	1/s
Packing density	—	s^2
Power density	—	s

Note: Analysis is based on the assumptions of on-chip loads with constant fan-out and constant clock frequency.

dissipate tens of watts per square centimeter. This level of dissipation presents special heat-removal problems, especially in the confines of portable computers. Modern microprocessors are thus fitted with miniature heat sinks and air-cooling fans; however, it is now necessary to scale down supply voltages in order to reign in the power density.

9.17 PSPICE Simulations

Simulations were performed using PSPICE 9.1, which can be downloaded free.[29] The MOSFET model parameters used in all simulations are provided in Table 9.6 and Table 9.7. The process transconductance parameters were calculated assuming an oxide thickness of 100 Å. For n-MOSFETS,

$$KP = \frac{(3.9)\left(8.85\times10^{-14}\ \mathrm{F/cm}\right)\left(580\ \mathrm{cm^2V^{-1}s^{-1}}\right)}{100\times10^{-8}\ \mathrm{cm}} = 0.20\ \mathrm{mA/V^2} \qquad (9.57)$$

and, for p-MOSFETS,

$$KP = \frac{(3.9)\left(8.85\times10^{-14}\ \mathrm{F/cm}\right)\left(230\ \mathrm{cm^2V^{-1}s^{-1}}\right)}{100\times10^{-8}\ \mathrm{cm}} = 0.079\ \mathrm{mA/V^2}\ . \qquad (9.58)$$

TABLE 9.6

n-MOSFET SPICE Parameters

Parameter	Value	Units
VTO	0.5	V
KP	0.2m	A/V^2
LAMBDA	0.05	—
CGSO	1.15n	F/m
CGDO	0.58n	F/m
VMAX	100k	m/s

TABLE 9.7

p-MOSFET SPICE Parameters

Parameter	Value	Units
VTO	–0.5	V
KP	0.079m	A/V^2
LAMBDA	0.05	—
CGSO	1.15n	F/m
CGDO	0.58n	F/m
VMAX	80k	m/s

The oxide capacitance per unit gate width was calculated assuming a gate length of 0.5 μm and an oxide thickness of 100 Å:

$$\frac{C_{OX}}{W} = \frac{(3.9)\left(8.85\times10^{-14}\,\text{F/cm}\right)\left(0.5\times10^{-4}\,\text{cm}\right)}{100\times10^{-8}\,\text{cm}} \tag{9.59}$$

$$= 17.3\ \text{pF/cm} = 1.73\ \text{nF/m}.$$

The individual contributions were estimated using

$$CGSO \approx \left(\frac{2}{3}\right)\left(\frac{C_{OX}}{W}\right) = 1.15\ \text{nF/m} \tag{9.60}$$

and

$$CGDO \approx \left(\frac{1}{3}\right)\left(\frac{C_{OX}}{W}\right) = 0.58\ \text{nF/m}. \tag{9.61}$$

9.17.1 Voltage Transfer Characteristic

The voltage transfer characteristic was simulated for the CMOS inverter of Figure 9.45 with V_{DD} = 2.5 V. The results are shown in Figure 9.46. The CMOS circuit exhibits rail-to-rail logic swing and a symmetric voltage transfer characteristic with V_{IL} = 1.04 V and V_{IH} = 1.46 V.

9.17.2 Short-Circuit Current

The short-circuit current was simulated for the CMOS inverter of Figure 9.47 with V_{DD} = 2.5 V. The results are shown in Figure 9.48 The peak supply current of 238 μA flows with $V_{IN} = V_{DD}/2$ = 1.25 V as expected.

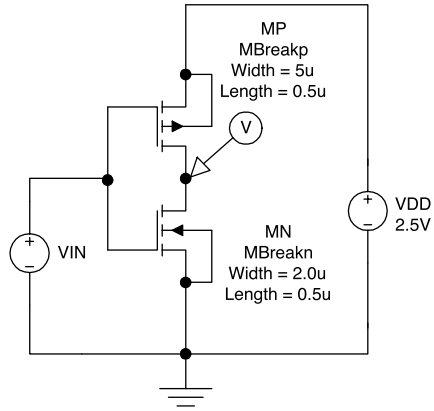


FIGURE 9.45
CMOS circuit used for simulation of the VTC.

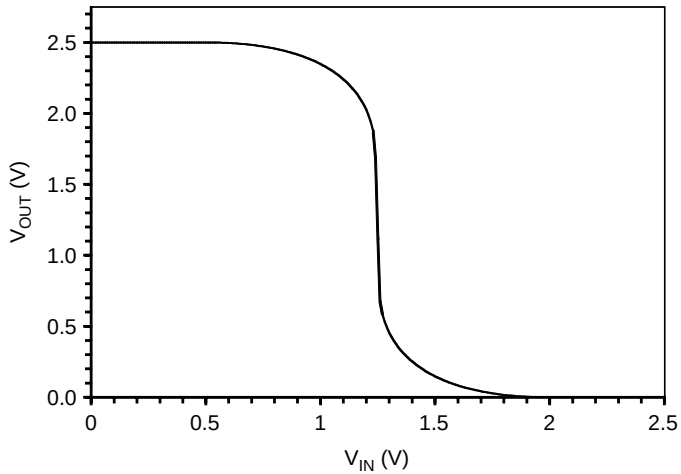


FIGURE 9.46
Simulated VTC for the CMOS circuit.

9.17.3 Propagation Delays

The propagation delays for the CMOS inverter with a 15-pF load were determined using the circuit shown in Figure 9.49. The pulse source parameters were V1 = 0 V, V2 = 2.5, TD = TR = TF = 0, PW = 100 ns, and PER = 200 ns. The results of the transient simulation appear in Figure 9.50. The propagation delays are equal, as expected for a symmetric CMOS inverter Additional transient simulations were done to determine the propagation delays as a function of the load capacitance; the results are shown in Figure 9.51. The propagation delay is directly proportional to the load capacitance and decreases monotonically with the supply voltage as expected.

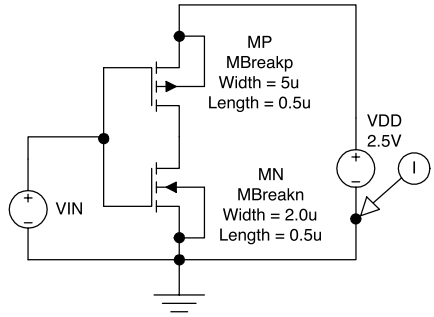


FIGURE 9.47
CMOS circuit used for simulation of the DC supply current.

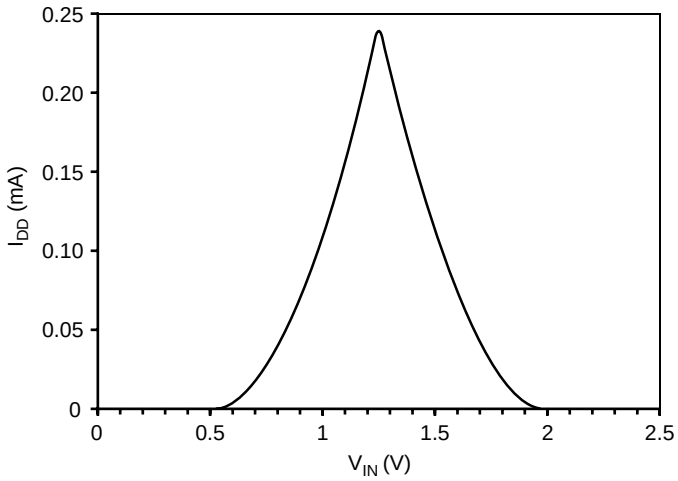


FIGURE 9.48
Simulated supply current vs. input voltage characteristic for the CMOS circuit.

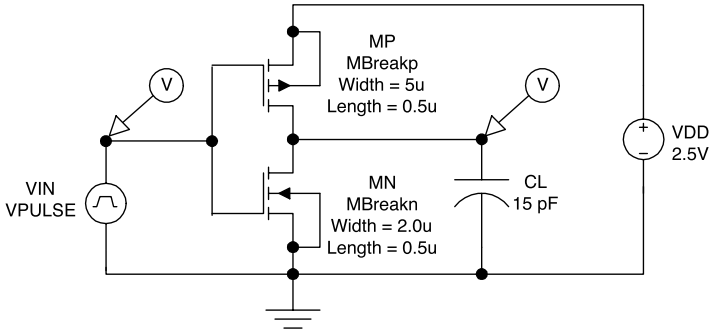


FIGURE 9.49
CMOS circuit used for simulation of the propagation delays.

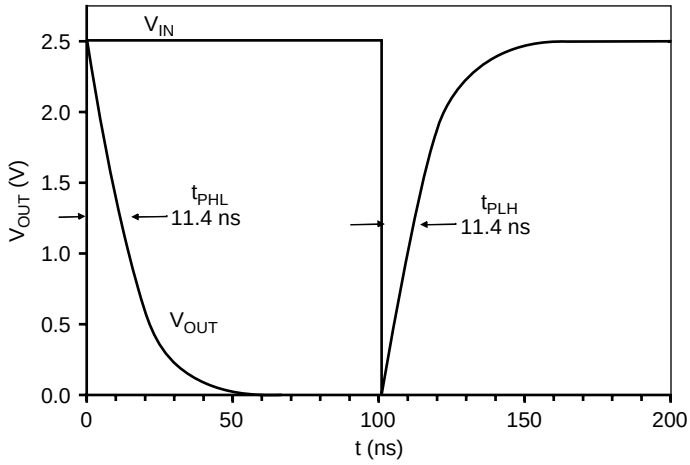


FIGURE 9.50
Simulated CMOS transient response.

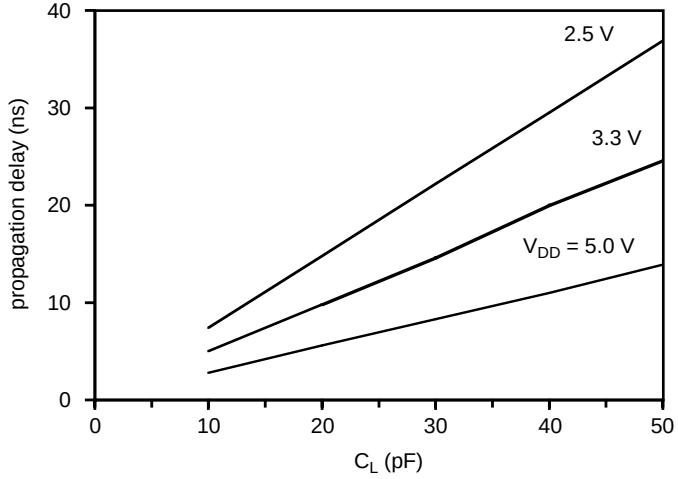


FIGURE 9.51
CMOS propagation delay as a function of the load capacitance with the supply voltage as a parameter.

9.17.4 Ring Oscillator

The transient response for a three-stage ring oscillator was simulated using the circuit of Figure 9.52; simulation results are shown in Figure 9.53. A 15-pF lumped capacitive load was placed at the output of each stage of the ring oscillator. The oscillations built up after approximately 0.5 μs without the need for external stimulus; the oscillation period was 83 ns. From this, the propagation delay may be calculated to be

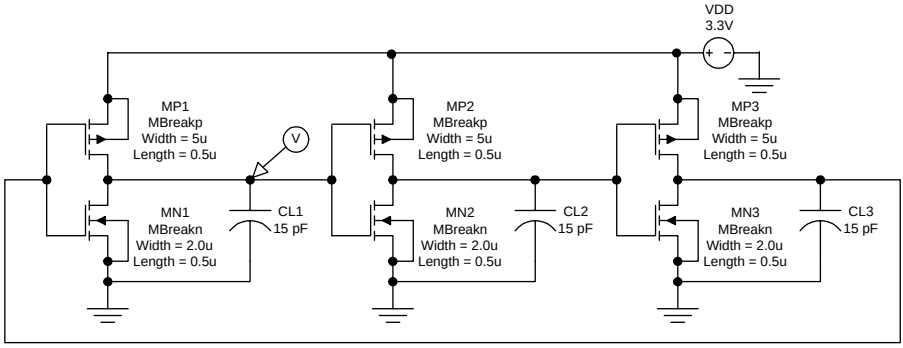


FIGURE 9.52
Three-stage CMOS ring oscillator with 15-pF loads at each stage.

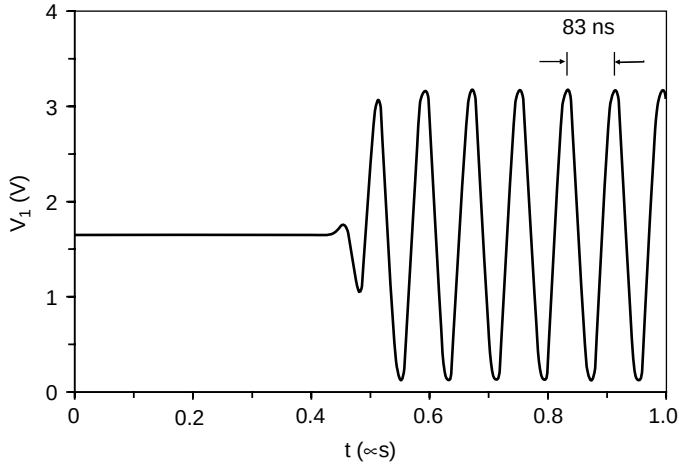


FIGURE 9.53
Transient response for three-stage CMOS ring oscillator.

$$t_P = \frac{83 \text{ ns}}{(2)(3)} = 13.8 \text{ ns}. \tag{9.62}$$

The propagation delay obtained using the ring oscillator is ~20% larger than the propagation delay using a square wave input. The difference is due to the approximately sinusoidal wave established at each stage of the ring oscillator.

9.17.5 Logic Function

The logic function of a CMOS NAND2 gate was investigated using a transient simulation for the circuit of Figure 9.54; results of the transient simulation

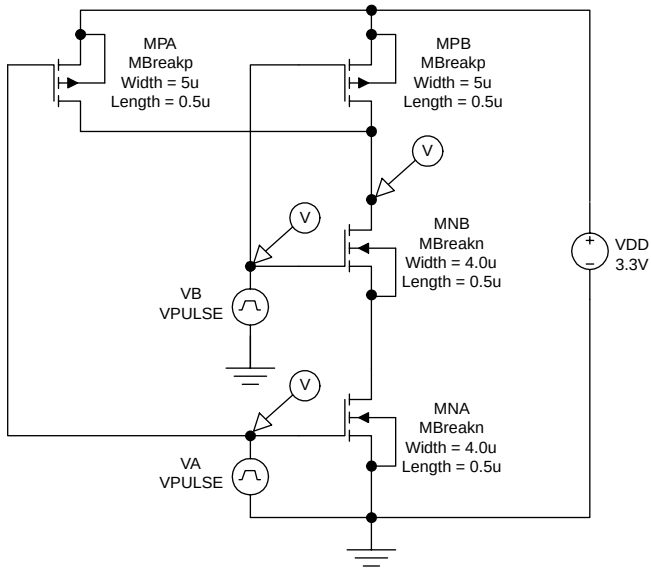


FIGURE 9.54
CMOS circuit used for the analysis of the logic function.

appear in Figure 9.55. The VB pulse source was set to twice the frequency of the VA pulse source. The results verify the expected NAND function for the circuit: the output node goes low if and only if logic one is applied to both inputs. Another notable feature of the transient response in Figure 9.55 is the dip in the output voltage at $t = 0.011$ s that appears as a result of charge sharing, as described in Section 9.12.

9.17.6 Dynamic CMOS

The behavior of a dynamic CMOS inverter was investigated using a transient simulation for the circuit illustrated in Figure 9.56; 50-pF lumped capacitive loads were placed at the two internal nodes of the dynamic inverter circuit. The clock was simulated using a pulse source with rise and fall times of 100 ps, a pulse width of 1 ns, and a period of 2.5 ns. The transient simulation was performed for the case in which the input waveform makes a low-to-high transition at $t = 3.0$ ns.

The waveforms for the clock, input, and output are provided in Figure 9.57. The first high-to-low transition of the clock turns on the p-MOSFET and precharges the output node to 3.3 V. During the first evaluation time interval (following the low-to-high transition of the clock), the input is low and the output node stays high at 3.3 V. The low-to-high transition at the input causes the output voltage to dip because of charge sharing. At the second evaluation interval (following the second low-to-high transition of the clock), the output goes low, as expected.

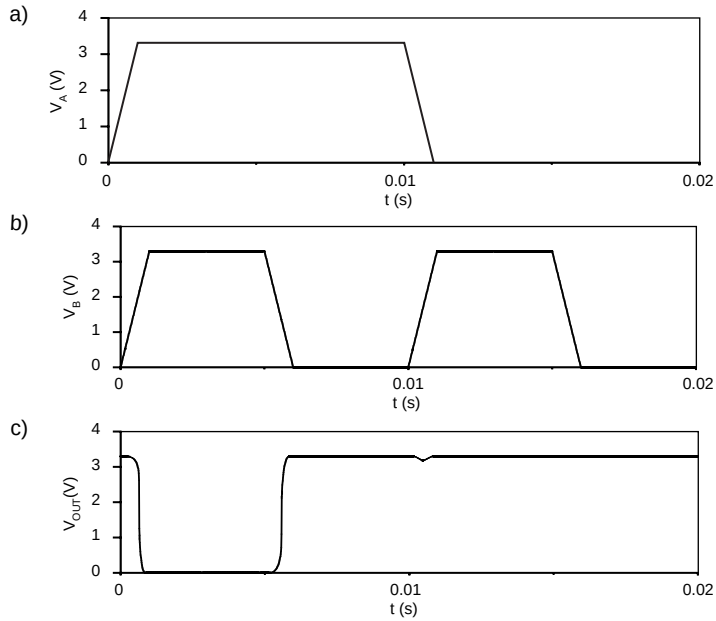


FIGURE 9.55
Simulated transient response for a CMOS NAND2 circuit.

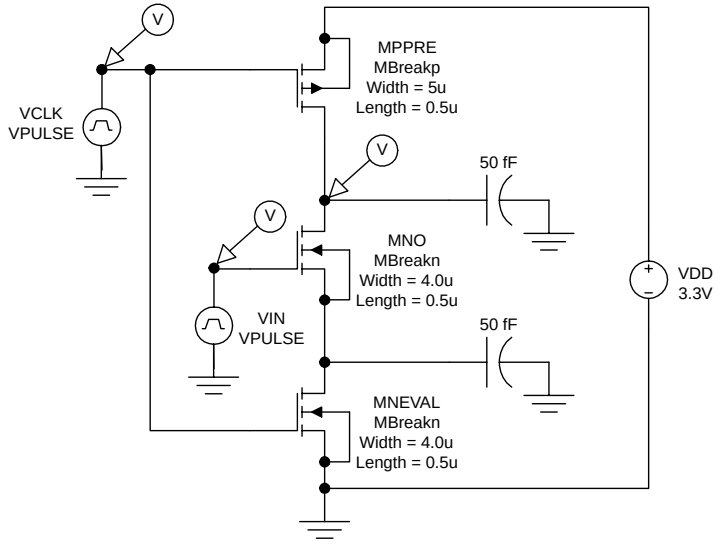


FIGURE 9.56
Dynamic CMOS circuit used for simulation of the transient response.

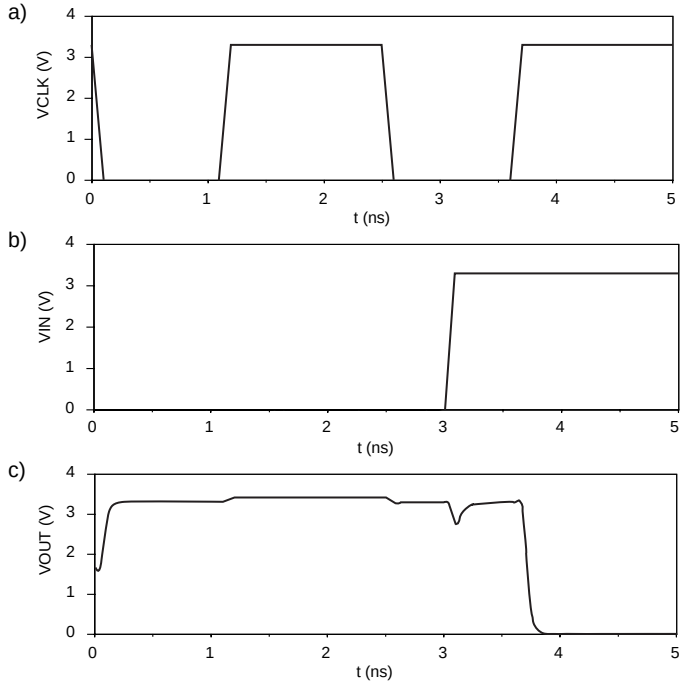


FIGURE 9.57
Simulated transient response for a dynamic CMOS inverter circuit.

9.18 Summary

Complementary metal oxide–semiconductor (CMOS) logic is the most important logic family today. CMOS gates, constructed using complementary pairs of n-MOSFETs and p-MOSFETs, exhibit nearly ideal voltage transfer characteristics, very low DC dissipation, high packing density, and high speed. The voltage transfer characteristic for a CMOS inverter can be calculated in five pieces corresponding to the different modes of operation for the two devices. For each regime of the characteristic, the output voltage can be determined by equating the drain currents in the n-channel and p-channel MOSFETs. For a symmetric CMOS inverter, the voltage transfer characteristic is nearly ideal, with rail-to-rail swing and a rather abrupt transition at one half of the supply voltage.

The dissipation in a CMOS gate is the sum of the static (DC) and dynamic (AC) components. The static dissipation is associated with the subthreshold currents in the cutoff MOSFETs and the leakage currents in the reverse-biased source and drain p–n junction diodes. The dynamic dissipation is the sum of the short-circuit power and the capacitance switching power. Often the

capacitance switching power is dominant in CMOS circuitry, so the dissipation increases linearly with the switching frequency.

For a symmetric CMOS gate, the propagation delay is directly proportional to the load capacitance and inversely proportional to the device transconductance parameters. The high-to-low and low-to-high propagation delays are equal because the transistors have matched device transconductance parameters. The maximum fan-out for a CMOS gate, loaded by similar gates, is determined entirely by dynamic considerations because the propagation delay increases linearly with the load capacitance and, therefore, the number of load gates.

The NAND function can be realized in CMOS by placing n-MOSFETs in series and p-MOSFETs in parallel. The NOR function is realized by placing the n-MOSFETs in parallel but the p-MOSFETs in series. Complex logic functions can be realized in CMOS by combining series and parallel branches of transistors. In such complex gates, the MOSFETs must be scaled appropriately to maintain acceptable DC and transient characteristics.

Dynamic, or clocked, CMOS gates achieve higher packing densities and lower dissipation than static CMOS gates. The principle underlying the operation of dynamic CMOS is that each circuit node exhibits capacitance and can be charged to a particular voltage level. Once charged, such a node can be read at a later time as long as the stored charge has not leaked away. Dynamic CMOS gates may be constructed using an NMOS pull-down network or a PMOS pull-up network. In either case, precharge and evaluation transistors are also included. Domino logic is an important family of dynamic CMOS with stages that alternate between the use of pull-up and pull-down networks.

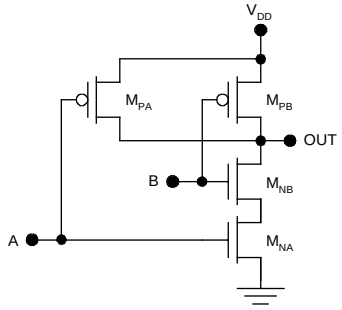
In integrated form, complementary pairs of n-MOSFETs and p-MOSFETs contain parasitic bipolar junction transistors that form two pnpn thyristors. Once turned on, either of these thyristors will latch on, thus causing excessive and destructive current flow. This problem, called *latch-up*, must be prevented by careful design of CMOS transistors. Using epitaxial layers and appropriate layout design rules have all but eliminated latch-up in modern CMOS circuits.

MOSFETs are susceptible to damage by static discharge, especially the scaled-down devices used in modern CMOS circuits. Static discharge involves the destructive breakdown of the gate oxide and must be avoided. In modern CMOS circuits, electrostatic discharge protection circuitry is provided at each pin; bipolar devices such as bipolar transistors or thyristors are used for this purpose. All the same, care must be taken in handling or assembling CMOS integrated circuits.

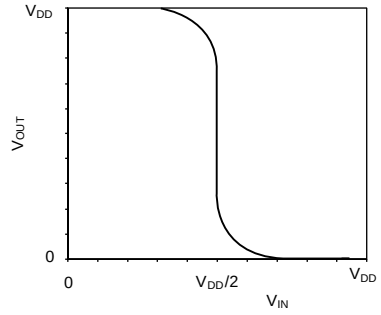
Scaling of CMOS devices has resulted in tremendous gains in circuit performance and packing density. Full scaling and constant voltage scaling are of particular interest, although some CMOS parameters such as V_T and t_{ox} cannot be scaled arbitrarily. The progress in this area has been exponential, as described by Moore's law.

CMOS LOGIC QUICK REFERENCE

NAND2 circuit



DC Voltage Transfer Characteristic



The high packing density and low power make CMOS the most important logic family today. The performance as well as the density of CMOS circuitry is improving rapidly with advances in lithography and the associated device scaling.

DC Voltage Transfer Characteristic

Range of V_{IN}	p-MOS	n-MOS	Equation for V_{OUT}
$V_{IN} \le V_T$	linear	cutoff	$V_{OUT} = V_{DD}$
$V_T \le V_{IN} \le V_{OUT} - V_T$	linear	sat	$V_{OUT} = (V_{IN} + V_T) + \sqrt{(V_{IN} - V_{DD} + V_T)^2 - (V_{IN} - V_T)^2}$
$V_{OUT} - V_T \le V_{IN} \le V_{OUT} + V_T$	sat	sat	(interpolate)
$V_{OUT} + V_T \le V_{IN} \le V_{DD} - V_T$	sat	linear	$V_{OUT} = (V_{IN} - V_T) - \sqrt{(V_{IN} - V_T)^2 - (V_{IN} - V_{DD} + V_T)^2}$
$V_{IN} \ge V_{DD} - V_T$	cutoff	linear	$V_{OUT} = 0$

Dissipation

$$P = \underbrace{P_{subthreshold} + P_{pn}}_{P_{DC}} + \underbrace{P_{sc} + P_{switch}}_{P_{AC}} \qquad P \approx P_{switch} \approx \alpha f_{CLK} C_L V_{DD}^2$$

Propagation Delays

$$t_{PLH} = t_{PHL} = t_P \approx \frac{C_L}{K}\left[\frac{2V_T}{(V_{DD} - V_T)^2} + \frac{2}{(V_{DD} - V_T)}\ln\left(\frac{V_{DD} - V_T}{V_{DD}/2}\right)\right]$$

Fan-out

$$N_{MAX} \le \frac{C_{L,\max}}{C_{IN}} \qquad C_{L,\max} \approx \frac{Kt_{P,\max}}{\left[\frac{2V_T}{(V_{DD} - V_T)^2} + \frac{2}{(V_{DD} - V_T)}\ln\left(\frac{V_{DD} - V_T}{V_{DD}/2}\right)\right]}$$

$$C_{IN} = C_{OXN} + C_{OXP} = \frac{\varepsilon_{OX} W_N L_N}{t_{OX}} + \frac{\varepsilon_{OX} W_P L_P}{t_{OX}}$$

Design Rules

$$\mu_n = 580 cm^2 V^{-1} s^{-1} \qquad \mu_p = 230 cm^2 V^{-1} s^{-1} \qquad \varepsilon_{ox} = 3.9\varepsilon_0 \approx 3.45 \times 10^{-14} F/cm$$

f = 10^{-15} p = 10^{-12} n = 10^{-9} μ = 10^{-6} m = 10^{-3} k = 10^{3} M = 10^{6} G = 10^{9}

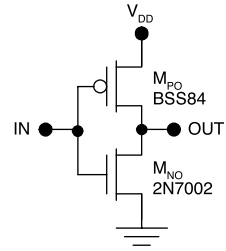


FIGURE 9.58
Unloaded CMOS inverter (L9.1).

Laboratory Exercises

L9.1. Consider the *unloaded* CMOS inverter of Figure 9.58 with 4 V ≤ V_{DD} ≤ 6 V. Obtain the data sheets for both types of transistors from the manufacturer's or vendor's Web sites.

1. Using hand calculations, determine and plot the voltage transfer characteristic for the case of V_{DD} = 5 V using the *nominal* device characteristics.
2. Using SPICE, determine and plot V_{OUT} vs. V_{IN} with V_{DD} as a parameter. (Three to five different values of V_{DD} are sufficient.)
3. Build the inverter and measure the voltage transfer characteristic using the x–y feature of an oscilloscope or virtual instrument, and a low frequency input signal (1 kHz).
4. Plot the hand-calculated, simulated, and experimental results on one graph of V_{OUT} vs. V_{IN} with V_{DD} as a parameter. Are there any significant differences among the three sets of results? Can you explain these differences?
5. Which of the critical voltages (V_{IL}, V_{IH}, V_{OL}, V_{OH}) are strongly affected by changes in V_{DD}? Which of the critical voltages are strongly affected by variations in the device threshold voltages?

L9.2. Consider the *unloaded* CMOS inverter illustrated in Figure 9.59. Obtain the data sheets for both types of transistors from the manufacturer's or vendor's Web sites.

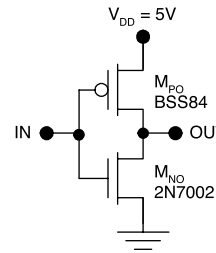


FIGURE 9.59
Unloaded CMOS inverter (L9.2).

1. Determine and plot the unloaded voltage transfer characteristic using hand calculations and the *nominal* device characteristics from the data sheets.
2. Using SPICE, determine and plot the unloaded voltage transfer characteristic using the *nominal* device characteristics from the data sheets.
3. Repeat the SPICE simulation of the VTC using the *minimum* value of V_{TN} and the *most negative* value of V_{TP}.
4. Repeat the SPICE simulation of the VTC using the *maximum* value of V_{TN} and the *least negative* value of V_{TP}.
5. Build the inverter circuit. Measure and plot the experimental VTC along with the hand-calculated and SPICE results.
6. Are there appreciable differences between the measured and hand-calculated voltage transfer characteristics? Can these be explained by the tolerances in the threshold voltages? Can you match the experimental result by adjusting only the threshold voltages in the SPICE models?

L9.3. Consider the *unloaded* CMOS inverter shown in Figure 9.60. Obtain the data sheets for both types of transistors from the manufacturer's or vendor's Web sites.

1. Determine and plot the unloaded I_{DD} vs. V_{IN} characteristic using hand calculations and the *nominal* device characteristics from the data sheets.
2. Using SPICE, determine and plot the unloaded I_{DD} vs. V_{IN} characteristic using the *nominal* device characteristics from the data sheets.
3. Using SPICE, determine the expected range of the peak supply current by considering the tolerances in the threshold voltages and the device transconductance parameters.
4. Build the inverter circuit. Measure and plot the experimental I_{DD} vs. V_{IN} characteristic along with the hand-calculated and SPICE results.

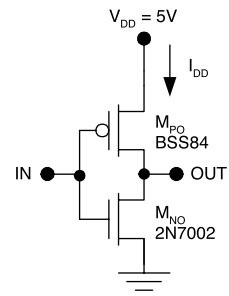


FIGURE 9.60
Unloaded CMOS inverter (L9.3).

5. What is the value of V_{IN} corresponding to the peak value of I_{DD}? Can you match the experimental value with SPICE by adjusting the threshold voltages in the SPICE models?
6. What is the peak value of I_{DD}? Can you match this experimental value by adjusting the threshold voltages and the device transconductance values in the SPICE models?

L9.4. Consider the CMOS inverter and NAND2 gate depicted in Figure 9.61. Obtain the data sheets for both types of transistors from the manufacturer's or vendor's Web sites.

1. Using hand calculations, estimate R_{OL} and R_{OH} for the inverter using the *nominal* device characteristics. (R_{OL} is the output resistance of the gate with a low (logic-zero) output and R_{OH} is the output resistance of the gate with a high (logic-one) output.
2. Using hand calculations, estimate R_{OL} and R_{OH} for the NAND2 gate using the *nominal* device characteristics. Is the NAND2 gate symmetric?
3. Using SPICE, estimate R_{OL} and R_{OH} for the inverter using the *nominal* device characteristics.
4. Using SPICE, estimate R_{OL} and R_{OH} for the NAND2 gate using the *nominal* device characteristics.
5. Build the inverter and experimentally determine R_{OL} and R_{OH}.
6. Build the NAND2 gate and experimentally determine R_{OL} and R_{OH}.
7. Are the gates symmetric? Why or why not?
8. Are hand-calculated and SPICE results in agreement with the experimental results? If not, is it possible to match the experimental results by modifying the SPICE models for the devices?

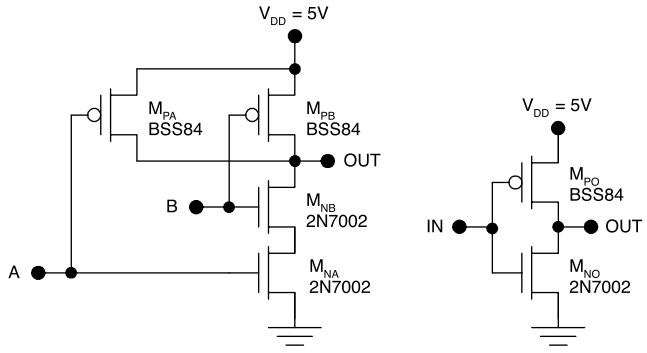


FIGURE 9.61
CMOS NAND2 gate and inverter (L9.4).

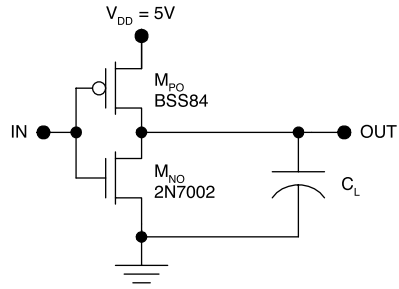


FIGURE 9.62
CMOS inverter (L9.5).

L9.5. Consider the CMOS inverter of Figure 9.62 with a lumped capacitive load. Obtain the data sheets for both types of transistors from the manufacturer's or vendor's Web sites.

1. Using hand calculations, determine and plot the propagation delay vs. the load capacitance.
2. Using SPICE transient simulations for a single CMOS inverter as shown previously, determine and plot the propagation delay vs. the load capacitance.
3. Build a three-stage CMOS ring oscillator using the circuit shown earlier. By varying the load capacitance (applied at each stage), determine and plot the average propagation delay vs. the load capacitance. Is the amplitude of the oscillations equal to the expected logic swing, V_{DD}?
4. Plot the hand-calculated, simulated, and experimental results together for direct comparison. Determine the slopes of the characteristics, in ohms. Are there significant differences (>10%) among the sets of data?
5. Use SPICE to predict the ring oscillator frequency vs. the load capacitance. Are these results in agreement with the transient results for a single CMOS gate? Are these results in agreement with the experimental results?

L9.6. Consider the CMOS inverter of Figure 9.63 with a lumped capacitive load. Obtain the data sheets for both types of transistors from the manufacturer's or vendor's Web sites.

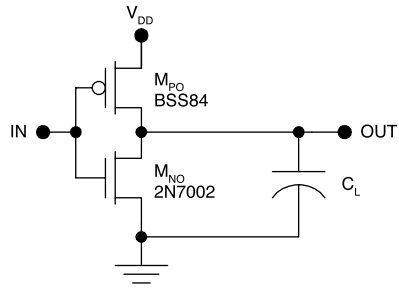


FIGURE 9.63
CMOS inverter (L9.6).

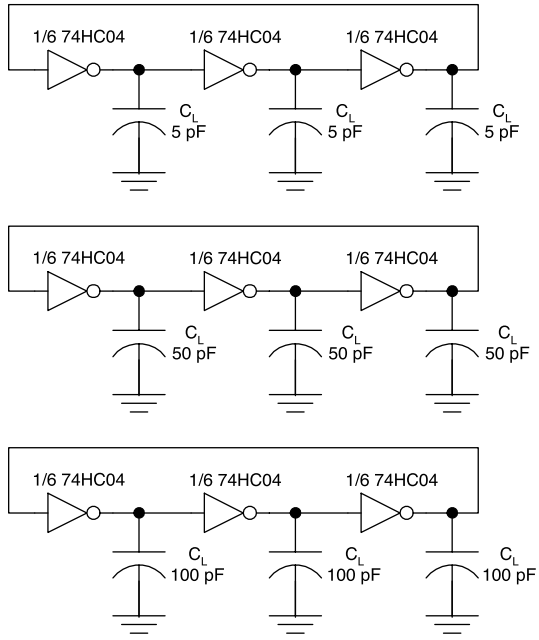


FIGURE 9.64
Three-stage ring oscillators (L9.7).

1. Using SPICE transient simulations for a three-stage CMOS ring oscillator, determine and plot t_P vs. C_L with V_{DD} as a parameter.
2. Using three-stage ring oscillators, experimentally determine t_P vs. C_L with V_{DD} as a parameter. Is the peak-to-peak amplitude for the oscillations equal to the expected logic swing? Is it a function of C_L?
3. Plot the SPICE and experimental results together for direct comparison. Determine the slopes of the characteristics in ohms. Is the slope a function of V_{DD}? Why or why not?
4. Can the experimental characteristics be fit by the equation $t_p = A\frac{C_L}{V_{DD}}$, where A is a constant?

L9.7. Consider three-stage ring oscillators built using 74HC CMOS inverters as shown in Figure 9.64.

1. Using hand calculations, estimate and plot the propagation delay vs. the load capacitance, with the supply voltage as a parameter.
2. Using SPICE, estimate and plot the propagation delay vs. the load capacitance, with the supply voltage as a parameter.
3. Build a seven-stage ring oscillator using seven of these inverter circuits. Based on the experimental results, plot the average

propagation delay vs. the load capacitance, with the supply voltage as a parameter.

4. Is it possible to fit the results with an expression of the form $t_p = A\frac{C_L}{V_{DD}}$, where A is a constant?

Problems

P9.1. Consider the CMOS gate shown in Figure 9.65. $K = 0.2$ mA/V^2 and $V_T = 0.6$ V.

1. Determine the mode of operation for each of the transistors.
2. Determine the supply current I_{DD}.
3. Determine the value of V_{OUT}.

P9.2. Consider the CMOS gate shown in Figure 9.66. $K = 0.22$ mA/V^2 and $V_T = 0.5$ V.

1. Determine the mode of operation for each of the transistors.
2. Determine the supply current I_{DD}.
3. Determine the value of V_{OUT}.

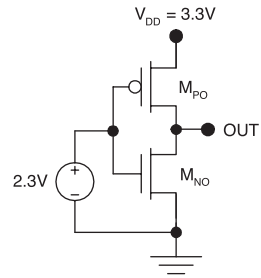


FIGURE 9.65
CMOS gate (P9.1).

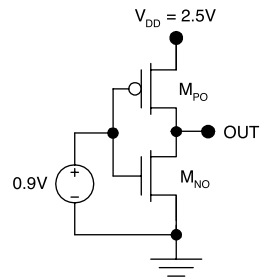


FIGURE 9.66
CMOS gate (P9.2).

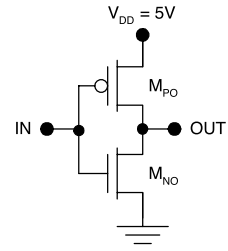


FIGURE 9.67
CMOS gate (P9.3).

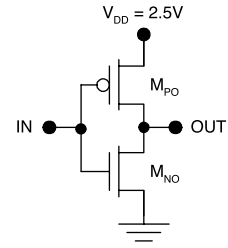


FIGURE 9.68
CMOS gate (P9.4).

P9.3. Consider the CMOS gate illustrated in Figure 9.67. $K = 0.15$ mA/V^2 and $V_T = 0.6$ V.

1. Using hand calculations, determine and plot the voltage transfer characteristic for the inverter.
2. Using the results of (1), determine V_{IL}, V_{IH}, V_{OL}, V_{OH}, and the noise margins.
3. Using hand calculations, determine and plot the supply current vs. V_{IN} for the inverter.
4. Determine the range of V_{IN} for which the n-MOSFET is linear.
5. Determine the range of V_{IN} for which the p-MOSFET is linear.

P9.4. Consider the CMOS gate illustrated in Figure 9.68. $K = 0.15$ mA/V^2 and $V_T = 0.45$ V.

1. Using hand calculations, determine and plot the voltage transfer characteristic for the inverter.
2. Using SPICE, determine and plot the voltage transfer characteristic for the inverter.
3. Using hand calculations, determine and plot the supply current vs. V_{IN} for the inverter.
4. Using SPICE, determine and plot the supply current vs. V_{IN} for the inverter.

P9.5. Consider the 74HC04 inverter illustrated in Figure 9.69.

1. Using hand calculations, determine and plot the voltage transfer functions for node B, node C, and the output node.
2. Repeat using SPICE.

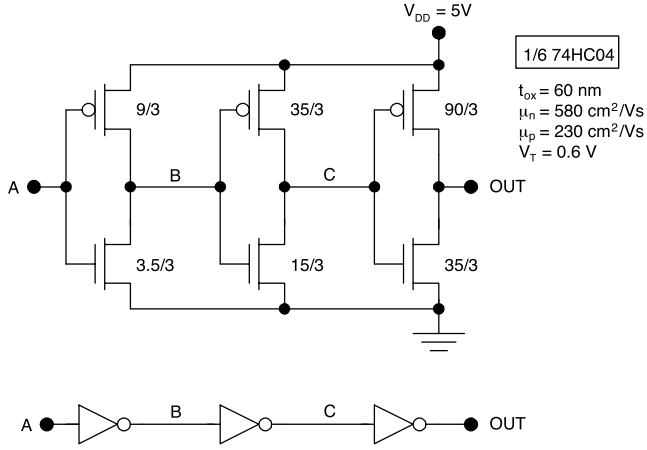


FIGURE 9.69
74HC04 inverter (P9.5).

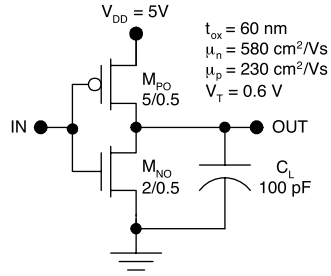


FIGURE 9.70
CMOS gate (P9.6).

P9.6. Consider the CMOS gate of Figure 9.70.
1. Using hand calculations, estimate the propagation delays.
2. Determine the propagation delays using SPICE. Show that t_{PLH} and t_{PHL} are equal, and compare the values to the hand calculation.

P9.7. For the CMOS gate shown in Figure 9.71, estimate the maximum fan-out if $t_{P,MAX}$ = 300 ps.

P9.8. For the CMOS gate of Figure 9.72, estimate the maximum fan-out if the clock frequency is to be 100 MHz.

P9.9. Consider CMOS circuitry driving a 50-pF off-chip load as shown in Figure 9.73. t_{ox} = 16 nm. V_T = 0.5 V. Suppose that each successive stage is scaled up in current driving capability by a factor of 3.
1. Determine the required number of buffer stages such that $t_P \leq$ 0.6 ns.
2. Use SPICE to verify your design.

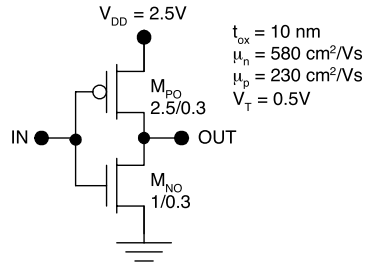


FIGURE 9.71
CMOS gate (P9.7).

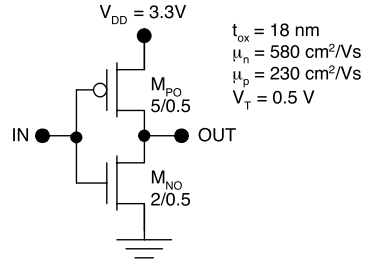


FIGURE 9.72
CMOS gate (P9.8).

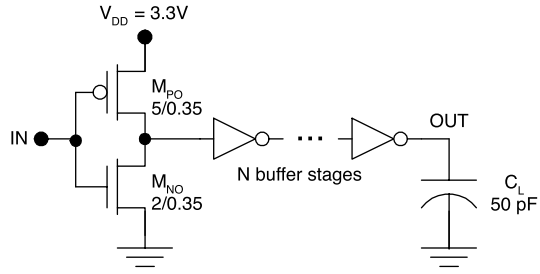


FIGURE 9.73
CMOS circuitry (P9.9).

P9.10. Consider CMOS circuitry driving a 50-pF off-chip load as illustrated in Figure 9.74. t_{ox} = 11 nm. V_T = 0.5 V.

1. Determine the minimum number of buffer stages necessary such that $t_P < 1$ ns.
2. For each buffer stage, determine the required gate dimensions for each of the MOSFETs.
3. Use SPICE to verify your design.

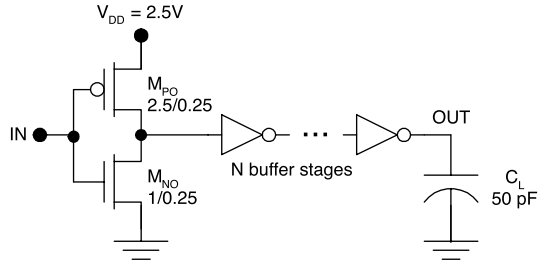


FIGURE 9.74
CMOS circuitry (P9.10).

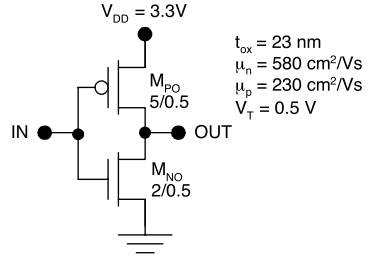


FIGURE 9.75
CMOS gate (P9.11).

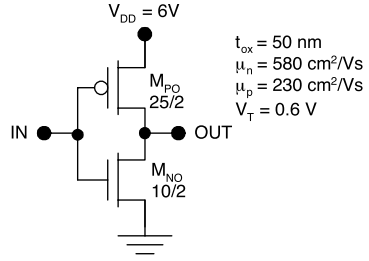


FIGURE 9.76
CMOS gate (P9.12).

P9.11. For the CMOS gate shown in Figure 9.75, perform an analysis of constant voltage scaling. (Assume that the oxide thickness and all gate dimensions are scaled by 1/s). $1 \leq s \leq 10$.

1. Determine and plot the propagation delay vs. s, assuming 20 on-chip loads.
2. Determine and plot the dissipation vs. s, assuming 20 on-chip loads and $f = 0.05/t_p$.

P9.12. For the CMOS gate of Figure 9.76, perform an analysis of "full" scaling. (Assume that the oxide thickness, all gate dimensions, and V_{DD} are scaled by 1/s.) $1 \leq s \leq 5$.

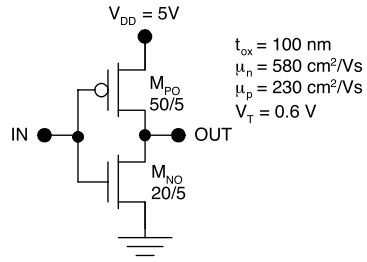


FIGURE 9.77
Gate representing 1975 technology (P9.13).

1. Determine and plot the propagation delay vs. s, assuming 20 on-chip loads.
2. Determine and plot the dissipation vs. s, assuming 20 on-chip loads and $f = 0.05/t_p$.

P9.13. The gate shown in Figure 9.77 represents 1975 technology. Assume that constant voltage scaling is used, with a scale factor of $1/\sqrt{2}$ applied every 3 years. Estimate and plot the microprocessor clock frequency vs. the year, up to the present time. Assume N = 20.

References

1. www.ti.com (Texas Instruments)
2. www.fairchildsemi.com (Fairchild Semiconductor)
3. www.toshiba.com (Toshiba)
4. Wakeman, L., DC electrical characteristics of MM74HC high-speed CMOS, Fairchild Semiconductor application note 313, www.fairchildsemi.com, 1998.
5. Allam, M.W., Anis, M.H., and Elmasry, M.I., High-speed dynamic logic styles for scaled-down CMOS and MTCMOS technologies, *Proc. 2000 Int. Symp. Low Power Electron. Design*, 155, 2000.
6. Srivastava, P., Pua, A., and Welch, L., Issues in the design of domino logic circuits, *Proc. 8th Great Lakes Symp. VLSI*, 108, 1998.
7. Wurtz, L., A scaling procedure for domino CMOS logic, *Proc. 1992 IEEE Southeast Conf.*, 2, 580, 1992.
8. Understanding latch-up in advanced CMOS logic, Fairchild Semiconductor application note 600, www.fairchildsemi.com, 1999.
9. Ker, M.-D., Lo, W.-Y., and Wu, C.-Y., New experimental methodology to extract compact layout rules for latch-up prevention in bulk CMOS ICs, *Proc. 1999 IEEE Custom Integrated Circuits*, 143, 1999.
10. Hargrove, M.J., Voldman, S., Gauthier, R., Brown, J., Duncan, K., and Craig, W., Latch-up in CMOS technology, *Proc. 36th IEEE Int. Reliability Phys. Symp.*, 269, 1998.
11. Amerasekera, A., Selvam, S.T., and Chapman, R.A., Designing latch-up robustness in a 0.35-µm technology, *Proc. 32nd IEEE Int. Reliability Phys. Symp.*, 280, 1994.

12. Ker, M.-D. and Chuang, C.-H., ESD implantations in 0.18-µm salicided CMOS technology for on-chip ESD protection with layout consideration, *Proc. 8th Int. Symp. Phys. Failure Anal. ICs*, 85, 2001.
13. Okushima, M., Noguchi, K., Sawahata, K., Suzuki, H., Kuroki, S., Koyama, S., Ando, K., and Ikezawa, N., ESD protection scheme using CMOS compatible vertical bipolar transistor for 130 nm CMOS generation, *Tech. Dig. 2000 Int. Electron. Devices Meet.*, 127, 2000.
14. Wang, T.-H. and Ker, M.-D., On-chip ESD protection design by using polysilicon diodes in CMOS technology for smart card application, *Proc. 2000 Electr. Overstress/Electrostatic Discharge Symp.*, 266, 2000.
15. Ker, M.-D., Whole-chip ESD protection design with efficient VDD-to-VSS ESD clamp circuits for submicron CMOS VLSI, *IEEE Trans. Electron. Devices*, 46, 173, 1999.
16. Lee, J., Syrzycki, M., and Iniewski, K., Semiconductor-controlled rectifier (SCR) electrostatic discharge (ESD) protection devices in submicron CMOS technology, *IEEE 1999 Can. Conf. Electr. Computer Eng.*, 1, 409, 1999.
17. Ker, M.-D., Chen, T.-Y., and Wu, C.-Y., CMOS on-chip ESD protection design with substrate-triggering technique, *IEEE 1998 Int. Conf. Electron. Circuits Syst.*, 1, 273, 1998.
18. Fried, R., Blecher, Y., and Friedman, S., CMOS ESD protection structures — characteristics and performance comparison, *Proc. 1995 Int. Semiconductor Conf.*, 567, 1995.
19. Yu, B., Wang, H., Xiang, O., An, J.X., Jeon, J., and Lin, M.-R., Scaling towards 35-nm gate length CMOS, *Dig. Tech. Pap. 2001 Symp. VLSI Technol.*, 9, 2001.
20. Shahidi, G.G., Challenges of CMOS scaling at below 0.1 µm, *Proc. 12th Int. Conf. Microelectron.*, 5, 2000.
21. Song, S., Yi, J.H., Kim, W.S., Lee, J.S., Fujihara, K., Kang, H.K., Moon, J.T., and Lee, M.Y., CMOS device scaling beyond 100 nm, *Tech. Dig. 2000 Int. Electron. Devices Meet.*, 235, 2000.
22. Taur, Y., The incredible shrinking transistor, *IEEE Spectrum*, 36, 25, 1999.
23. Timp, G., Bourdelle, K.K., Bower, J.E., Baumann, F.H., Boone, T., Cirelli, R., Evans-Lutterodt, K., Garno, J., Ghetti, A., Gossmann, H., Green, M., Jacobson, D., Kim, Y., Kleiman, R., Klemens, F., Kornlit, A., Lochstampfor, C., Mansfield, W., and Moccio, S., Progress toward 10-nm CMOS devices, *Tech. Dig. 1998 Int. Electron. Devices Meet.*, 615, 1998.
24. Chen, K., Hu, C., Fang, P., Gupta, A., Lin, M.P., and Wollesen, D., Experimental and analytical studies on CMOS scaling in deep submicron regime including quantum and polysilicon gate depletion effects, *Dig. 55th Device Res. Conf.*, 20, 1997.
25. Davari, B., CMOS technology scaling, 0.1 µm and beyond, Proc. 1996 Int. Electron Devices Meet., 555, 1996.
26. Davari, B., Dennard, R.H., and Shahidi, G.G., CMOS scaling for high performance and low power — the next 10 years, *Proc. IEEE*, 83, 595, 1995.
27. Chen, Z., Shott, J., Burr, J., and Plummer, J.D., CMOS technology scaling for low-voltage, low-power applications, *Dig. Tech. Pap. 1994 IEEE Symp. Low Power Electron.*, 56, 1994.
28. Taur, Y. and Mi, Y.-J., 0.1 _m CMOS and beyond, *Proc. 1993 Int. Symp. VLSI Technol., Syst., Appl.*, 1, 1993.
29. www.cadence.com (Cadence)

10

Low-Power CMOS Logic

10.1 Introduction

Low-power CMOS[1] has become increasingly important because of the proliferation of portable and hand-held electronic products. In these applications, battery lifetime is of critical importance. Over the past two decades, the battery energy density (in joules per kilogram) has roughly doubled whereas microprocessor dissipation has increased 50-fold. A second issue is heat removal. High-power integrated circuits present difficult challenges in packaging and heat sinking, adding cost, size, and weight to the products in which they are used.

The most effective way to reduce the power in CMOS circuitry is to scale down the supply voltage, but this involves a trade-off with speed. Sometimes multiple supply voltages are used so that critical path circuitry can use higher supply voltages to optimize speed.[2] However, fixed supply voltages must be chosen for worst-case throughput conditions. Another approach is dynamic voltage scaling (DVS), in which the supply voltage is adjusted dynamically to provide only the required throughput and therefore minimum dissipation.[3–8]

In low-voltage circuits, it is necessary to scale down the threshold voltages to maintain reasonable dynamic performance. However, this is accompanied by higher subthreshold conduction. As a practical rule of thumb, acceptable subthreshold leakage dictates that the threshold voltages should be at least two to three times the subthreshold swing.[8,9] However, this restriction can be lifted by using variable threshold CMOS (active body biasing)[10–14] or multiple threshold CMOS.[15,16] Another solution to this problem is the use of silicon-on-insulator (SOI).[17–44] SOI devices have superior subthreshold swings, allowing the operation of SOI CMOS with V_T = 0.1 V and V_{DD} = 0.4 V at room temperature.

Other approaches to power reduction in CMOS involve reduction of the switching activity or the load capacitances, or charge recycling. The switching activity can be reduced by clock-gating or sleep-control techniques. Reducing load capacitances can be achieved by scaling the devices; in addition,

a higher level of integration can minimize the need for driving off-chip load capacitances. As a rule of thumb, combining the functionality of four chips into one can cut the overall power dissipation by half. Charge recycling (reusing electrical charges for more than one logic operation) is achieved in adiabatic logic circuits,[44–47] but these require a different circuit topology than that used in conventional CMOS or domino logic.

In this chapter, circuit approaches to low-power CMOS design will be described, with an emphasis on the principles involved. System-level approaches to low power design are also increasingly important, but beyond the scope of this book. The reader is referred to books on VLSI design for coverage of this topic.

10.2 Low-Voltage CMOS

As described in Chapter 9, the dissipation in a CMOS gate is given by

$$P = \underbrace{P_{subthreshold} + P_{leakage}}_{P_{DC}} + \underbrace{P_{sc} + P_{switch}}_{P_{AC}}, \tag{10.1}$$

where

$P_{subthreshold}$ = power associated with MOSFET subthreshold conduction
$P_{leakage}$ = power associated with p–n junction leakage in the MOSFETs
P_{SC} = short-circuit dissipation
P_{switch} = capacitance switching dissipation

In high-voltage CMOS circuits, the capacitance switching power is nearly always dominant so that

$$P \approx P_{switch} = \alpha f_{CLK} C_L V_{DD}^2, \tag{10.2}$$

where V_{DD} is the supply voltage, C_L is the load capacitance, f_{CLK} is the clock frequency, and α is the switching activity factor. Aggressive scaling of the physical dimensions of the MOSFETs has led to significant reductions in the load capacitances and the dissipation per gate. Beyond that, the most effective way to reduce power is to decrease the supply voltage; however, reducing the supply voltage increases the propagation delays, given by

$$t_P \approx \frac{C_L}{K}\left[\frac{2V_T}{(V_{DD}-V_T)^2} + \frac{2}{(V_{DD}-V_T)}\ln\left(\frac{V_{DD}-V_T}{V_{DD}/2}\right)\right] \approx \frac{1.6C_L}{K(V_{DD}-V_T)}, \tag{10.3}$$

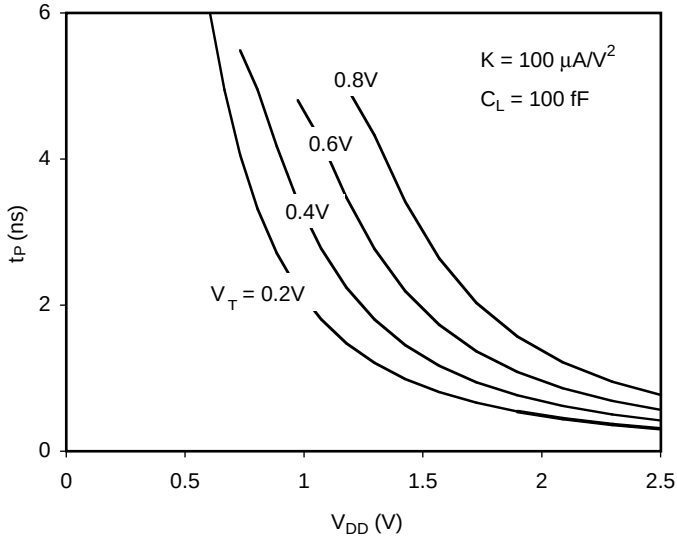


FIGURE 10.1
Propagation delay vs. supply voltage with threshold voltage as a parameter for a CMOS gate with K = 100 μA/V^2 and loaded by 100 fF.

where K is the device transconductance parameter for the MOSFETs and V_T is the absolute value of the threshold voltages in the MOSFETs. This equation shows that scaling the supply voltage should be accompanied by reducing the threshold voltages to maintain reasonable dynamic performance.

Figure 10.1 shows the propagation delay vs. the supply voltage with the threshold voltage as a parameter for a symmetric CMOS gate (K = 100 μA/V^2) loaded by 100 fF. This figure shows that reduction of the threshold voltage is an effective way to maintain the dynamic performance while reducing the supply voltage (and therefore dissipation).

On the other hand, reducing the threshold voltages increases the subthreshold currents and the associated dissipation. The subthreshold dissipation is

$$P_{subthreshold} \approx V_{DD}K(1+m)\left(\frac{kT}{q}\right)^2 \exp\left(\frac{-V_T}{mkT/q}\right), \tag{10.4}$$

where k is the Boltzmann constant, T is the absolute temperature, q is the electronic charge, and m ranges between one and two for conventional MOSFETs operated at room temperature. This corresponds to a subthreshold swing between 60 and 120 mV. In fixed threshold CMOS circuits, the absolute value of the threshold voltages must be at least two to three times the subthreshold swing to obtain acceptable subthreshold conduction.[1] With a typical subthreshold swing of 100 mV, CMOS circuitry operating at 300 K can be designed with V_{DD} = 1.0 V and V_T = 0.3 V.

10.3 Multiple-Voltage CMOS

A problem that arises in low-voltage CMOS is the degradation of the switching speed for the output buffer drivers. If acceptable off-chip data rates are to be maintained using a single, low supply voltage, the output drivers must be made very wide and take up considerable chip area. This problem can be alleviated by using multiple supply voltages because then circuits with more critical speed requirements can operate at higher voltages while other circuitry can operate at lower voltages to minimize power. An added benefit of this approach is that the off-chip signals have increased voltage swing and, thus, noise margins.

If many different supply voltages are available, then all circuits can operate with just sufficient throughput to avoid wasting power. However, this increases the complexity of the power supply circuitry and its distribution network, so a trade-off occurs between power-performance and complexity. Therefore, it is common to use only two different supply voltages: a high voltage for output drivers and a low supply voltage for all internal circuitry.

Example 10.1

Suppose that a CMOS integrated circuit must drive 15-pF off-chip loads with a maximum propagation delay of 1 ns. Assuming 0.25-μm technology with t_{ox} = 10 nm, determine the minimum widths of the output driver transistors assuming V_T = 0.3 V and V_{DD} = 5 V. Repeat for the case of V_{DD} = 1 V.

Solution. For output drivers operating at 5 V, the device transconductance parameters must be at least

$$K \geq \frac{C_L}{t_{P,\max}}\left[\frac{2V_T}{(V_{DD}-V_T)^2}+\frac{2}{(V_{DD}-V_T)}\ln\left(\frac{V_{DD}-V_T}{V_{DD}/2}\right)\right]$$

$$=\left(\frac{15\times10^{-12}\text{ F}}{10^{-9}s}\right)\left[\frac{2(0.3\text{ V})}{(5\text{ V}-0.3\text{ V})^2}+\frac{2}{(5V-0.3\text{ V})}\ln\left(\frac{5\text{ V}-0.3\text{ V}}{5\text{ V}/2}\right)\right].$$

$$=4.4\times10^{-3}\text{ A}/\text{V}^2$$

The process transconductance parameters are

$$k'_N=\frac{\mu_n\varepsilon_{ox}}{t_{ox}}=\frac{(580\text{ cm}^2\text{V}^{-1}\text{s}^{-1})(3.9)(8.85\times10^{-14}\text{ F/cm})}{10\times10^{-7}\text{ cm}}=200\ \mu\text{A}/\text{V}^2$$

and

$$k'_P = \frac{\mu_p \varepsilon_{ox}}{t_{ox}} = \frac{\left(230\ \text{cm}^2\text{V}^{-1}\text{s}^{-1}\right)(3.9)\left(8.85\times10^{-14}\ \text{F/cm}\right)}{10\times10^{-7}\ \text{cm}} = 80\ \mu\text{A/V}^2\ ;$$

therefore, the required widths for the output drivers are

$$W_N = \left(\frac{4400\ \mu\text{A/V}^2}{200\ \mu\text{A/V}^2}\right)0.25\ \mu\text{m} = 5.5\ \mu\text{m}$$

and

$$W_P = \left(\frac{4400\ \mu\text{A/V}^2}{80\ \mu\text{A/V}^2}\right)0.25\ \mu\text{m} = 13.8\ \mu\text{m}\,.$$

However, if the output drivers operate at 1 V, then the device transconductance parameters must be at least

$$\begin{aligned} K &\geq \frac{C_L}{t_{P,\max}}\left[\frac{2V_T}{\left(V_{DD}-V_T\right)^2} + \frac{2}{\left(V_{DD}-V_T\right)}\ln\left(\frac{V_{DD}-V_T}{V_{DD}/2}\right)\right] \\ &= \left(\frac{15\times10^{-12}\ \text{F}}{10^{-9}s}\right)\left[\frac{2(0.3\ \text{V})}{(1\ \text{V}-0.3\ \text{V})^2} + \frac{2}{(1\ \text{V}-0.3\ \text{V})}\ln\left(\frac{1\ \text{V}-0.3\ \text{V}}{1\ \text{V}/2}\right)\right]; \\ &= 3.3\times10^{-2}\ \text{A/V}^2 \end{aligned}$$

therefore, the required widths for the output drivers are

$$W_N = \left(\frac{3.3\times10^{-2}\ \text{A/V}^2}{200\ \mu\text{A/V}^2}\right)0.25\ \mu\text{m} = 41\ \mu\text{m}$$

and

$$W_P = \left(\frac{3.3\times10^{-2}\ \text{A/V}^2}{80\ \mu\text{A/V}^2}\right)0.25\ \mu\text{m} = 102\ \mu\text{m}\,.$$

Therefore, the 1-V output drivers will require 7.5 times as much chip area as the 5-V drivers. The need for buffer stages will make the situation even worse.

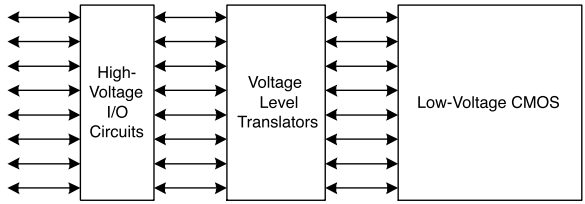


FIGURE 10.2
Dual-voltage CMOS.

In dual-voltage CMOS, the internal (low-voltage) circuitry can be optimized for low power while the (high-voltage) output drivers can operate at higher voltage to conserve die area. All circuits are designed to just meet the speed requirements only enough to avoid wasting power or die area. Because the I/O circuits operate at a higher voltage than the internal circuitry, it is necessary to introduce voltage level shifters (Figure 10.2).

10.4 Dynamic Voltage Scaling

The choice of supply voltage always involves a trade-off between speed and power, regardless of the choice of threshold voltage. In CMOS circuits operating with a single supply voltage, it is necessary to choose the supply voltage to meet the speed requirements in the most critical circuitry — often the I/O circuits. This results in wasted power in other parts of the integrated circuit.

The use of two or more supply voltages alleviates this problem to a great extent because the output drivers can operate at a high supply voltage, for high speed, while the core of the integrated circuit can operate at a lower supply voltage for reduced dissipation. Nevertheless, the speed requirements vary significantly within the core of the circuitry, from circuit to circuit and also over time. Therefore, the choice of any single supply voltage for the core of the circuitry will result in wasted power. Dynamic voltage scaling (DVS) allows the supply voltage to be adjusted dynamically for each block of circuitry. By making the supply voltage just adequate for the required throughput (speed), the power can be minimized.

The two basic approaches to DVS are (1) to use a finite number of discrete supply voltages (discrete V_{DD} scaling) and (2) to use a continuous variation of the supply voltage (arbitrary V_{DD} scaling). In either case, the supply voltage is produced by a switching power supply driven by a feedback control system.

Typically, DVS is used in conjunction with a variable clock frequency. Thus the clock frequency and the supply voltage can be adjusted as appropriate for the necessary throughput in order to optimize the dissipation, as realized

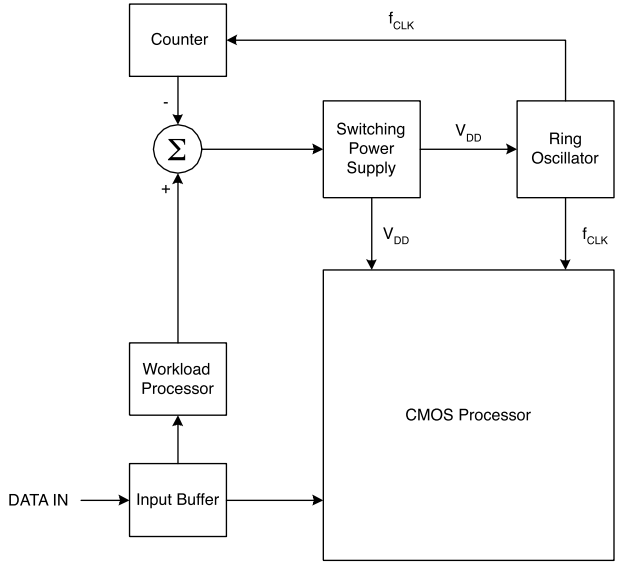


FIGURE 10.3
Dynamic voltage scaling (DVS) system.

in Figure 10.3. Here, the continuously adjustable V_{DD} is provided by the switching power supply that feeds the CMOS processor and a ring oscillator that mimics the critical path in the CMOS processor. The workload processor determines the minimum required clock frequency based on the throughput requirement, as estimated by sampling the input. The feedback control of the switching power supply adjusts the V_{DD} until it is sufficient so that the ring oscillator operates at a frequency determined by the workload processor. Then the CMOS processor operates at the minimum necessary values of V_{DD} and f_{CLK} to provide the required throughput, thus minimizing the capacitance switching dissipation.

The control circuitry associated with DVS is relatively complex. As such, DVS is only practical if used with large blocks of circuitry.

10.5 Active Body Biasing

In low-voltage CMOS, choice of the threshold voltage involves a trade-off between the speed and the subthreshold conduction. The active biasing scheme avoids this trade-off because the threshold voltages of the transistors are adjusted dynamically. Therefore, circuits designed in this way are often called variable threshold CMOS (VT CMOS). The threshold adjustment is achieved through the body bias effect. A substrate bias control circuit drives the bodies of the n-MOSFETs and p-MOSFETs for this purpose. Active biasing

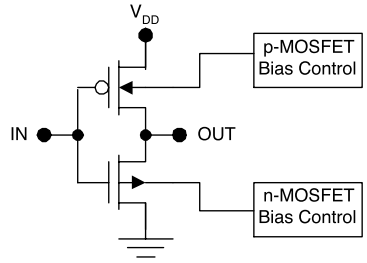


FIGURE 10.4
CMOS inverter with active body biasing.

requires an increase in the use of silicon area, to support the bias control circuitry. Despite this, the approach is commonly employed in modern microprocessor designs.

The basis for active biasing is the body effect. If a non-zero bias V_{BS} is applied between the source and the body of an n-MOSFET, then the threshold voltage is modified to

$$V_T = V_{TO} + \gamma\left(\sqrt{|V_{BS} + 2\phi_F|} - \sqrt{|2\phi_F|}\right), \tag{10.5}$$

where

V_{TO} = zero-bias threshold voltage
$2\phi_F$ = voltage across the semiconductor necessary to create a conducting channel (inversion layer)
γ = body effect coefficient

In an n-MOSFET, the threshold voltage can be made more positive by applying a negative bias to the body with respect to the source. In a p-MOSFET, the threshold voltage can be made more negative by applying a positive bias on the body. A general active biasing scheme takes the form shown in Figure 10.4. Separate bias control networks drive the n-MOSFETs and p-MOSFETs.

When the circuitry is actively switching, the source-to-body voltages are made zero. In other words, the bodies of the n-MOSFETs are biased at 0 V while the bodies of the p-MOSFETs are biased at V_{DD}. This results in normal operation of the circuitry, with the nominal threshold voltages as determined by the fabrication process. During standby operation, a negative bias is applied to the body of the n-MOSFET and a positive bias is applied to the body of the p-MOSFET. Therefore, both threshold voltages are increased in absolute value, greatly reducing the subthreshold conduction.

Using active biasing allows the use of variable threshold voltages, thus removing the trade-off between speed and subthreshold conduction. Therefore, the nominal threshold voltages can be much lower than those used in fixed threshold circuits. This in turn facilitates the reduction of the supply voltage for low-power operation.

An added benefit of active biasing is that it can be used to correct for process-induced threshold voltage variations. In fixed threshold CMOS circuits, the minimum practical threshold voltage is determined in part by process tolerances. Correction for these process variations allows further reduction in the nominal threshold voltages without adverse consequences.

Example 10.2

Consider variable threshold CMOS with t_{OX} = 7.5 nm. The n-channel and p-channel MOSFETs have channel doping equal to 10^{16} cm^{-3}. Calculate the body biases necessary to change the threshold voltages from ±0.1 V to ±0.3 V.

Solution. The body effect coefficient for the n-channel MOSFETs is

$$\gamma_N = \frac{\sqrt{2q\varepsilon_{Si}N_A}}{C_{OX}/A}$$

$$= \frac{\sqrt{2\left(1.602\times10^{-19}\text{ C}\right)(11.9)\left(8.85\times10^{-14}\text{ F/cm}\right)\left(10^{16}\text{ cm}^{-3}\right)}}{(3.9)\left(8.85\times10^{-14}\text{ F/cm}\right)/7.5\times10^{-7}\text{ cm}}$$

$$= 0.126\text{ V}^{1/2};$$

also,

$$\phi_{FN} = \frac{kT}{q}\ln\left(\frac{n_i}{N_A}\right) = (0.026\text{ V})\ln\left(\frac{1.45\times10^{10}\text{ cm}^{-3}}{10^{16}\text{ cm}^{-3}}\right) = -0.35\text{ V}.$$

The necessary body bias is

$$V_{BSN} = -\left[\frac{\Delta V_T}{\gamma_N} + \sqrt{|2\phi_{FN}|}\right]^2 - 2\phi_{FN}$$

$$= -\left[\frac{0.2\text{ V}}{0.126\text{ V}} + \sqrt{0.70\text{ V}}\right]^2 + 0.70\text{ V} = -5.2\text{ V}.$$

For the p-channel MOSFETs,

$$\gamma_P = -\frac{\sqrt{2q\varepsilon_{Si}N_D}}{C_{OX}/A}$$

$$= -\frac{\sqrt{2\left(1.602\times10^{-19}\text{ C}\right)(11.9)\left(8.85\times10^{-14}\text{ F/cm}\right)\left(10^{16}\text{ cm}^{-3}\right)}}{(3.9)\left(8.85\times10^{-14}\text{ F/cm}\right)/7.5\times10^{-7}\text{ cm}}$$

$$= -0.126\text{ V}^{1/2};$$

also,

$$\phi_{FP} = \frac{kT}{q}\ln\left(\frac{N_D}{n_i}\right) = (0.026\text{ V})\ln\left(\frac{10^{16}\text{ cm}^{-3}}{1.45\times10^{16}\text{ cm}^{-3}}\right) = 0.35\text{ V}.$$

The necessary body bias is

$$V_{BSP} = \left[\frac{\Delta V_T}{\gamma_P} + \sqrt{|2\phi_{FP}|}\right]^2 - 2\phi_{FP}$$

$$= \left[\frac{-0.2\text{ V}}{-0.126\text{ V}} + \sqrt{0.70\text{ V}}\right]^2 - 0.70\text{ V} = +5.2\text{ V}.$$

Unfortunately, the required body bias voltages are large in comparison to the supply voltages used in low-power CMOS.

Example 10.3

Consider variable threshold CMOS with K = 1 mA/V² and V_{DD} = 1.0 V. Calculate the standby power as a function of V_T, assuming that the subthreshold conduction is dominant. T = 300 K and m = 1.6.

Solution. The standby power is given by

$$P_{subthreshold} \approx V_{DD}K(1+m)\left(\frac{kT}{q}\right)^2\exp\left(\frac{-V_T}{mkT/q}\right)$$

$$= (1.0\text{ V})\left(10^{-3}\text{ A/V}^2\right)(1+1.6)\left(26\times10^{-3}\text{ V}\right)\exp\left(\frac{-V_T}{42\text{ mV}}\right)$$

$$= \left(68\times10^{-6}\text{ W}\right)\exp\left(\frac{-V_T}{42\text{ mV}}\right).$$

The results are plotted in Figure 10.5.

10.6 Multiple Threshold CMOS

Multiple threshold CMOS (MT CMOS) circuits can be used to overcome the trade-off between speed and subthreshold conduction inherent in single threshold CMOS. Two or more distinct threshold voltages have been used

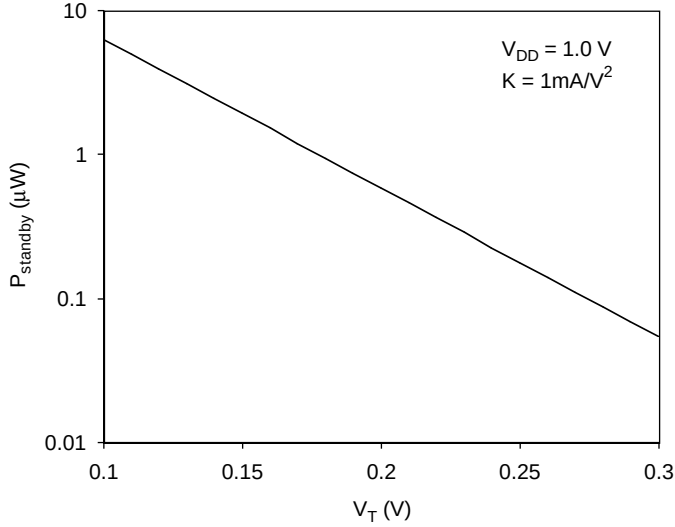


FIGURE 10.5
Standby dissipation vs. the threshold voltage for variable voltage CMOS with V_{DD} = 1.0 V and K = 1 mA/V².

in commercial microprocessors. The simplest multiple threshold scheme is dual threshold CMOS (DT CMOS), in which the logic circuits are disconnected from the supply rails during standby operation. Therefore, these logic circuits can use low threshold voltages for optimum speed. High threshold MOSFETs are used to disconnect the logic circuits from the supply rails. These devices have low subthreshold leakage, so the standby power can be greatly reduced by this approach.

Figure 10.6 illustrates the dual threshold concept. When the enable signal is high, the power-switching MOSFETs, M_{PS} and M_{NS}, are on, connecting the virtual V_{DD} line and virtual GND line to the power rails. In the standby state, the enable signal is brought low. Both power switching transistors are cut off in the standby state.

In dual threshold CMOS, the low threshold logic circuits are optimized for speed. The power-switching transistors are scaled up sufficiently so that they introduce less than a 10% increase in the propagation delays; however, the standby dissipation is determined by the high threshold power-switching transistors. These transistors are designed with large threshold voltages to reduce subthreshold conduction.

The application of DT CMOS circuits reduces the standby dissipation (compared to the case of low threshold CMOS) by a factor

$$\frac{P_{DTCMOS}}{P_{LTCMOS}} = \frac{K_H}{nK_L} 10^{-(V_{TH}-V_{TL})/S}, \tag{10.6}$$

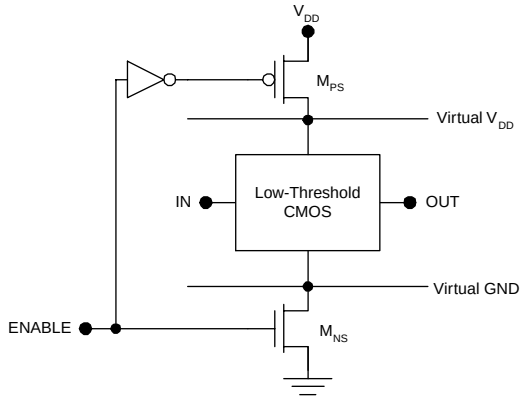


FIGURE 10.6
Dual-threshold CMOS.

where

S = subthreshold swing
n = number of low threshold logic gates served by one pair of power-switching transistors
K_H = device transconductance parameter for high threshold MOSFETs
K_L = device transconductance parameter for low threshold MOSFETs
V_{TH} = threshold voltage for high threshold MOSFETs
V_{TL} = threshold voltage for low threshold MOSFETs

A difficulty with dual threshold CMOS is that the circuits lose data when they are disconnected from the supply rails. To address this, special latches called balloon circuits are used to retain data in the "sleep" state.

Example 10.4

Estimate the reduction in the standby power associated with the use of DT CMOS with V_{TH} = 0.3 V and V_{TL} = 0.1 V.

Solution. It is assumed that the power-switching transistors are scaled sufficiently so that they increase the propagation delays of the low threshold circuits only by about 10%. Then

$$\frac{K_H}{nK_L} \approx 10 .$$

If it is assumed that the subthreshold swing is 100 mV, the reduction in the standby dissipation compared to the use of low threshold CMOS is approximately

$$\frac{P_{DTCMOS}}{P_{LTCMOS}} = \frac{K_H}{nK_L} 10^{-(V_{TH}-V_{TL})/S}$$

$$= (10)10^{-(0.3\text{ V}-0.1\text{ V})/0.1\text{ V}} = 0.1,$$

corresponding to savings of 90% in standby dissipation.

10.7 Adiabatic Logic

Adiabatic logic circuits[45–47] conserve power by recycling electrical charge, whereas conventional CMOS use each electrical charge only once. To see why this is so, consider the conventional CMOS circuit as shown in Figure 10.7 and Figure 10.8. Here, the load capacitance is charged through the p-MOSFET and discharges through the n-MOSFET. During the low-to-high transition shown in Figure 10.7, the load capacitance charges to V_{DD} and the energy $C_L V_{DD}^2/2$ is stored in the capacitor. An equal amount of energy is dissipated in the p-MOSFET. During the high-to-low transition illustrated in Figure 10.8, the energy stored in the capacitor is dissipated in the n-MOSFET and the stored charge is conducted to ground. Therefore, in conventional CMOS, the energy $C_L V_{DD}^2$ is dissipated for each complete switching cycle and each electrical charge is used only once.

The idea behind adiabatic logic is that charge may be reused to perform logic functions, without drawing additional charge from the power supply. Ideally, each charge would be recycled an infinite number of times. Although this ideal cannot be achieved, it is possible to reduce the dissipation significantly by recycling charge in real adiabatic circuits.

The concept of adiabatic switching can be understood with the aid of Figure 10.9 and Figure 10.10. In Figure 10.9, the load capacitance is charged through a p-MOSFET using a constant current source rather than a constant voltage source. During the charge-up of the load, the p-MOSFET is assumed to operate in the linear region of operation and is modeled using a resistance.

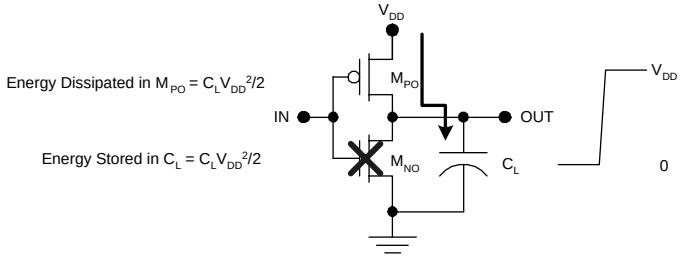


FIGURE 10.7
Charging of the load capacitance in conventional CMOS.

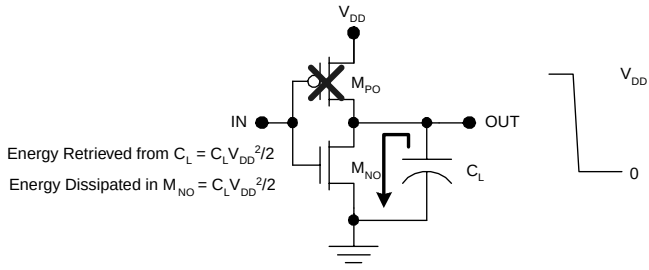


FIGURE 10.8
Discharge of the load capacitance in conventional CMOS.

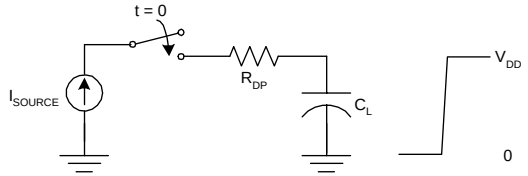


FIGURE 10.9
Charge-up of a load capacitance using a current source.

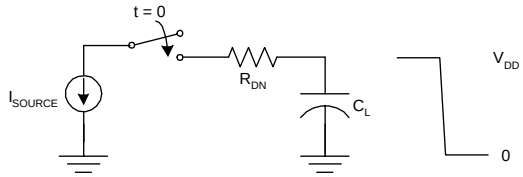


FIGURE 10.10
Discharge of the load capacitance using a current source.

If the initial voltage on the load capacitance is zero, then during the charge-up process, the voltage on the capacitor is given by

$$V_C(t) = \frac{I_{SOURCE}}{C_L} t \ . \tag{10.7}$$

The time required to charge the load up to V_{DD} is therefore

$$t_R = \frac{V_{DD} C_L}{I_{SOURCE}} \ . \tag{10.8}$$

During the charge-up process, the dissipation in the p-MOSFET is constant:

$$P_R = R_{DP} I_{SOURCE}^2, \tag{10.9}$$

so the energy dissipated in the p-MOSFET during the entire process of charging the capacitor from zero to V_{DD} is

$$J_R = V_{DD} C_L R_{DP} I_{SOURCE}. \tag{10.10}$$

Therefore, the energy wasted in the p-MOSFET may be made arbitrarily small if the rise time is made arbitrarily long.

Using a current source to discharge the load capacitance provides a similar benefit, as can be shown with the aid of Figure 10.10. If the initial voltage on the load capacitance is V_{DD}, then during the discharge process, the voltage on the capacitor is given by

$$V_C(t) = V_{DD} - \frac{I_{SOURCE}}{C_L} t. \tag{10.11}$$

The time required to discharge the load fully is therefore

$$t_F = \frac{V_{DD} C_L}{I_{SOURCE}}. \tag{10.12}$$

During the discharge process, the dissipation in the n-MOSFET is constant:

$$P_F = R_{DN} I_{SOURCE}^2. \tag{10.13}$$

Therefore, the energy dissipated in the n-MOSFET during the entire process of discharging the load capacitor from V_{DD} to zero is

$$J_F = V_{DD} C_L R_{DN} I_{SOURCE}. \tag{10.14}$$

and most of the energy initially stored in the capacitor can be returned to the supply for recycling, as long as the fall time is made long.

If the total energy wasted per switching cycle in the adiabatic circuit is compared to that in the conventional CMOS case, assuming symmetric circuits with equal fall and rise times ($t_F = t_R = \tau$), then

$$\frac{J_{Adiabatic}}{J_{Conventional}} = \frac{2 V_{DD} C_L R_D I_{SOURCE}}{V_{DD}^2 C_L} = \frac{2 V_{DD} C_L^2 R_D}{\tau I_{SOURCE}}. \tag{10.15}$$

In the adiabatic limit, this ratio will be zero. In real circuits, which dissipate a finite amount of power, a trade-off between the dissipation and the speed takes place.

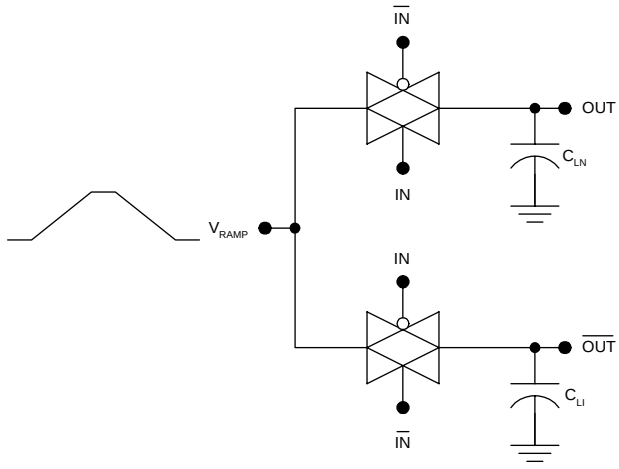


FIGURE 10.11
Adiabatic CMOS inverter–buffer.

In practical adiabatic circuits,* the current sources are implemented approximately using linear voltage ramps. Such a circuit is shown in Figure 10.11, which is an adiabatic inverter–buffer. With a logic-zero input, the bottom transmission gate turns on but the top transmission gate stays off. During the ramp-up of V_{RAMP}, the load capacitance C_{LI} at the inverting output charges with an approximately constant current, drawing energy from the ramp supply; the noninverting output stays at zero. Both of the complementary outputs may be evaluated at the end of the ramp-up. During the ramp-down process, the load capacitance at the inverting output discharges with an almost constant current. Unlike in conventional CMOS, the load capacitance discharges to the ramp supply rather than ground. During the ramp-down, therefore, most of the energy that had been stored in the load capacitance is returned to the ramp supply for use by other logic circuits.

In practice, the linear ramps in the supply voltage are approximated by a series of steps. This can be done by sequentially switching a number of fixed supply voltages through a number of n-MOSFETs. The use of N such voltage steps reduces the capacitance switching power by a factor of 1/N. Thus, adiabatic logic configured with 10 or fewer voltage steps can provide considerable savings in the switching dissipation.

More complex logic functions can be realized using adiabatic logic gates as well. The two-input AND/NAND gate is shown in Figure 10.12. It should be noted that each CMOS transmission gate requires two MOSFETs; therefore an adiabatic logic gate requires four MOSFETs per input, or double the number required by for conventional CMOS. This is because complementary

* These circuits are not *adiabatic* in the true sense of the word. They draw a reduced, but nonzero, amount of energy from the power supply.

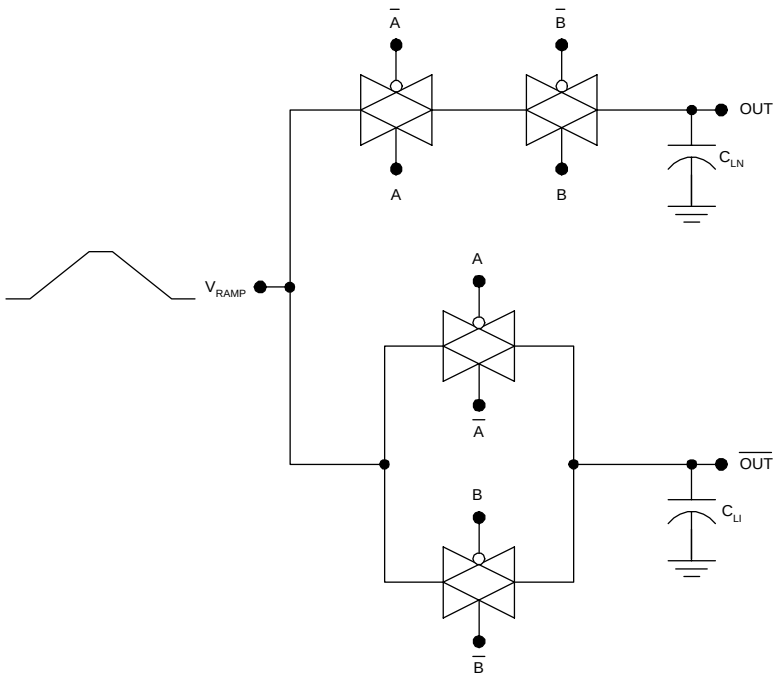


FIGURE 10.12
Adiabatic CMOS AND2/NAND2 gate.

signals must be provided to drive the fan-out gates. Thus, as with many of the other low-power CMOS strategies, a trade-off occurs between dissipation and circuit area.

10.8 Silicon-on-Insulator (SOI)

Silicon-on-insulator refers to any technology capable of producing silicon devices on an insulating substrate. It was originally conceived as a way to reduce device parasitics and improve the radiation hardness of silicon circuitry. However, it is now recognized as an important technology for low-power CMOS due to the improved subthreshold characteristics of SOI transistors compared to bulk silicon MOSFETs.

Over the years a number of SOI fabrication technologies have emerged. For example, silicon-on-sapphire (SOS) has been explored extensively for power devices. At the present time, the mainstream SOI technologies rely on silicon starting wafers, which are readily available with large area, high quality, and low cost. These technologies have made possible the implementation of commercial microprocessors and digital memories using SOI.

The following sections review two SOI fabrication technologies (**s**eparation by **im**plantation of **ox**ygen, or SIMOX, and wafer bonding, or WB). Then the features of fully depleted (FD) and partially depleted (PD) SOI transistors will be delineated. Finally, the application of SOI to low-power CMOS will be considered.

10.8.1 SOI Technologies: SIMOX and Wafer Bonding

At the present time there are two basic fabrication technologies for silicon-on-insulator: **s**eparation by **im**plantation of **ox**ygen (SIMOX)[32,33] and wafer bonding (WB).[34–37] In the SIMOX process, oxygen is implanted into a bulk silicon wafer, thus creating a buried oxide (BOX) layer. The devices are then fabricated in the thin silicon layer above the buried oxide. The wafer-bonding process involves bonding an oxidized device wafer (DW) to a handle wafer (HW), followed by removal of all but a thin layer from the device wafer. Two important variations of the WB process are the epitaxial layer transfer (ELTRAN) process and the UNIBOND™ process. SIMOX and WB approaches enjoy the advantage of being based on the same silicon wafers used for the fabrication of bulk CMOS devices.

The SIMOX process is illustrated in Figure 10.13 through Figure 10.15. First, a low dose (4×10^{17} cm^{-2}) of oxygen ions is implanted at an energy of about 200 keV (Figure 10.13). Next, the wafer is treated at a high temperature (>1300°C) to anneal out the defects created by the ion implantation process (Figure 10.14). Finally, the wafer is subjected to an internal thermal oxidation process (ITOX), which increases the thickness of the BOX layer to a usable value (Figure 10.15). Typically, the resulting SOI layer is 50 to 100 nm thick whereas the BOX layer is approximately 100 nm thick. Sometimes wafers produced by this method are called ITOX–SIMOX wafers.

The ELTRAN process is illustrated in Figure 10.16 through Figure 10.21. In contrast to the SIMOX process, wafer-bonding processes such as ELTRAN require two wafers, called the device wafer and the handle wafer.

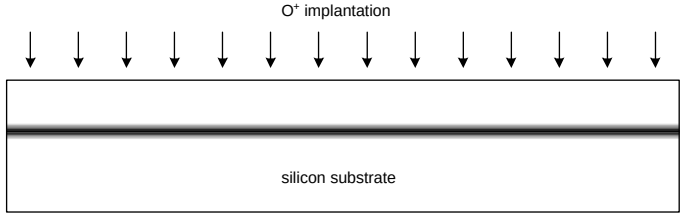


FIGURE 10.13
SIMOX process. The wafer is implanted with a low dose (4×10^{17} cm^{-2}) of oxygen at an energy of approximately 200 keV.

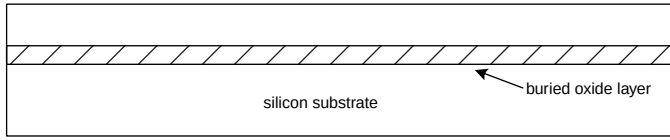


FIGURE 10.14
SIMOX process. Ion implantation damage is annealed out at >1300°C.

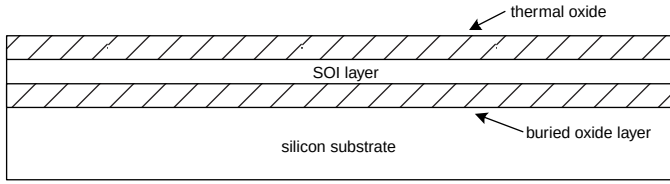


FIGURE 10.15
SIMOX process. Internal thermal oxidation (ITOX) at >1300°C is used to increase the thickness of the buried oxide (BOX) layer.

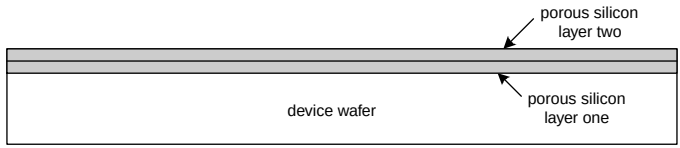


FIGURE 10.16
ELTRAN process. The device wafer is anodized in a two-step process to result in two layers of porous silicon with distinctly different porosities.

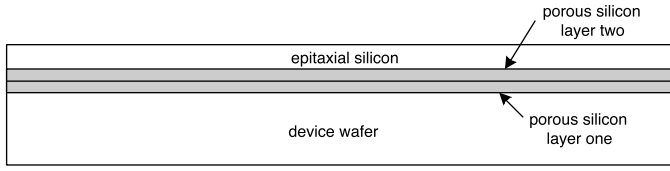


FIGURE 10.17
ELTRAN process. An epitaxial layer of device-quality silicon is grown on the porous silicon by vapor phase epitaxy (VPE).

1. The device wafer is treated by anodization to create a layer of porous silicon on its surface (Figure 10.16); the anodization process is designed to result in two distinct layers with different porosity.
2. A high-quality epitaxial layer of silicon is grown on top of the porous silicon by vapor phase epitaxy (Figure 10.17).

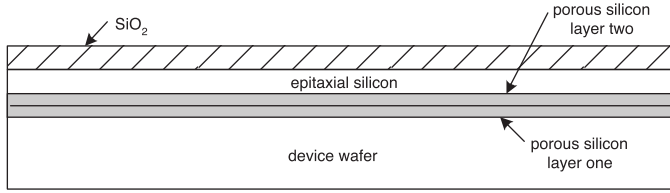


FIGURE 10.18
ELTRAN process. Thermal oxidation results in an oxide layer on top of the epitaxial layer of silicon.

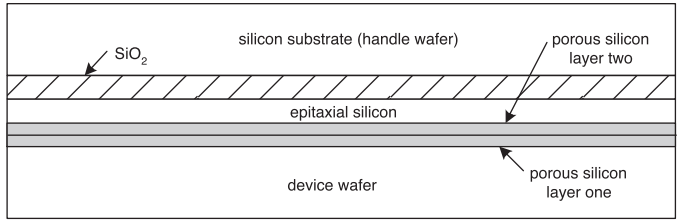


FIGURE 10.19
ELTRAN process. The handle wafer is bonded to the processed device wafer.

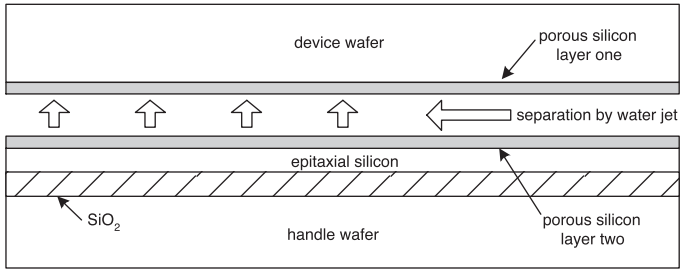


FIGURE 10.20
ELTRAN process. The handle and device wafers are separated between the distinct porous silicon layers by a water jet.

3. Thermal oxidation is used to create the oxide layer on top of the epitaxial device layer (Figure 10.18).
4. A handle wafer is bonded to the processed device wafer (Figure 10.19).
5. A water jet is used to separate the structure between the two distinct layers of porous silicon (Figure 10.20).
6. A chemical etch is used to remove the remaining porous silicon (Figure 10.21).

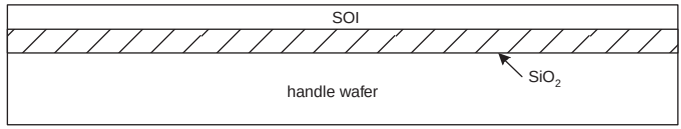


FIGURE 10.21
ELTRAN process. Chemical etching is used to remove the remaining porous silicon from the SOI layer.

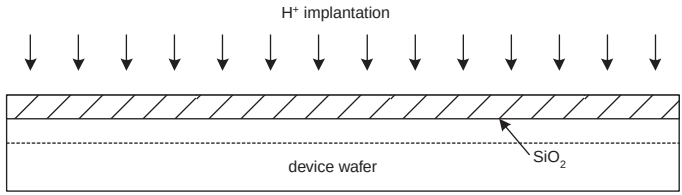


FIGURE 10.22
UNIBOND™ process. The device wafer is thermally oxidized and then implanted with hydrogen ions through the oxide.

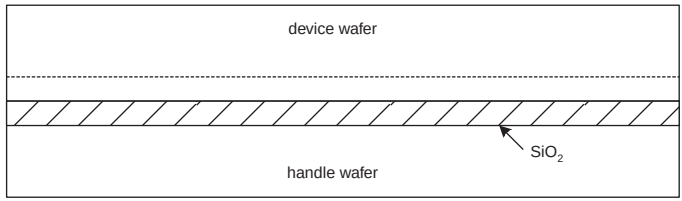


FIGURE 10.23
UNIBOND™ process. The device wafer is bonded to the handle wafer.

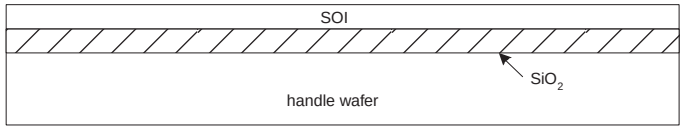


FIGURE 10.24
UNIBOND™ process. An annealing process at 400 to 600°C splits the wafers apart and the SOI wafer is chemically polished.

Another wafer bonding approach is the UNIBOND™ process, illustrated in Figure 10.22 through Figure 10.24. First, the device wafer is thermally oxidized. Then hydrogen ions are implanted through this oxide layer with a dose of about 10^{16} cm^{-2} (Figure 10.22). Next, the handle wafer is bonded to the device wafer (Figure 10.23). Then, an annealing process at 400 to 600°C serves to split the wafers apart and the resulting SOI wafer is chemically polished (Figure 10.24).

Many other variations on the basic SIMOX and WB processes can be envisioned. However, the important point is that these basic technologies can provide large-area, high-quality, and low-cost SOI wafers. These characteristics have made SOI commercially important for digital integrated circuits at the present time.

10.8.2 SOI MOSFETs: Fully Depleted or Partially Depleted

Depending on the thickness of the SOI layer, the bodies of the resulting MOSFETs may be partially depleted (PD)[38–40] or fully depleted (FD).[41–44] FD and PD SOI n-MOSFETs are depicted in Figure 10.25. Typically, FD MOSFETs are fabricated with an SOI layer approximately 50 nm thick. Therefore, with zero bias, the p-type silicon region under the channel becomes fully depleted by the built-in potentials at the source–drain junctions and the oxide interface. PD MOSFETs are fabricated using a thicker SOI layer, approximately 100 nm thick. Therefore an undepleted p-type body region exists under the channel of the PD SOI MOSFET at zero bias. PD and FD SOI MOSFETs have particular advantages and as such both device types have been applied in commercial microprocessors since 1999.

Partially depleted transistors are easier to manufacture because of the thicker layers and the less stringent process tolerances. This means that higher yields can be achieved using PD transistors. Nonetheless, FD MOSFETs have recently entered the mainstream of the microprocessor industry.

Active body biasing can be employed with partially depleted SOI MOSFETs to reduce subthreshold conduction or to compensate for processing variations in the threshold voltages. This is not possible with fully depleted devices because of the absence of a p-type body region. On the other hand, FD transistors exhibit near ideal (~60 mV at room temperature) values for the subthreshold swing; consequently, active body biasing is less important for these devices. In addition, the use of body contacts for the PD transistors consumes silicon area and reduces the packing density.

Partially depleted devices can be made with floating bodies (without body contacts). These devices exhibit a *floating body effect*, in which holes created by impact ionization near the drain accumulate in the body region. This

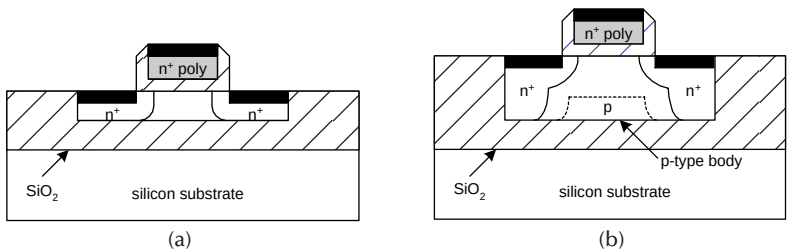


FIGURE 10.25
Fully depleted (FD) SOI n-MOSFET (a) and partially depleted (PD) SOI n-MOSFET (b).

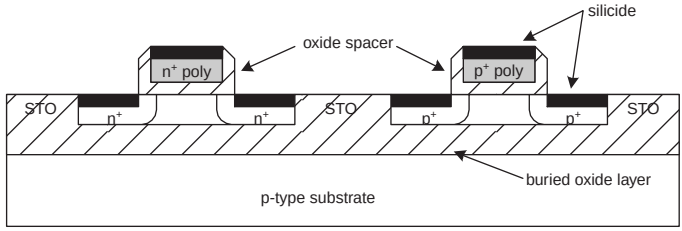


FIGURE 10.26
SOI CMOS transistors fabricated by the ITOX–SIMOX process.

creates a self-body bias that lowers the threshold voltage for the device, which can be advantageous in some logic circuits, where it increases the speed. In pass transistor applications such as dynamic random access memories (DRAMs), however, the floating body effect is problematic because the reduction in the threshold voltage increases the off-state leakage.

In the balance, it seems that FD SOI transistors may be more suitable for very low-voltage applications, whereas PD devices may be preferred in some higher-voltage, high-performance applications.

10.8.3 SOI for Low-Power CMOS

Silicon-on-insulator (SOI) is an important technology for low-power CMOS (see Figure 10.26) because SOI MOSFETs have superior subthreshold characteristics compared to bulk silicon MOSFETs. In FD SOI MOSFETs, the subthreshold swing is near the ideal value of 60 mV at room temperature. Combined with reduced process induced variations in the threshold voltage, this allows the implementation of SOI CMOS scaled down to V_{DD} = 0.4 V and V_T = 0.1 V. This reduction in the supply voltage decreases the capacitance switching power to one sixth of the value for bulk CMOS designed with V_{DD} = 1.0 V and V_T = 0.3 V. Further benefit derives from the reduction of the parasitic drain capacitances in SOI CMOS, which decreases the load capacitances (and therefore the switching dissipation) by about 20%.

The subthreshold swing for a MOSFET is given by

$$S = 2.3\frac{mkT}{q}, \tag{10.16}$$

where k is the Boltzmann constant, T is the absolute temperature, q is the electronic charge, and

$$m = 1 + \frac{C_{dm}}{C_{ox}}, \tag{10.17}$$

where C_{dm} is the depletion layer capacitance in the body of the MOSFET under inversion and C_{OX} is the gate oxide capacitance.

In fully depleted SOI MOSFETS such as those shown in Figure 10.25, the bodies of the devices deplete all the way through to the underlying oxide layer. Although the n-MOSFET is fabricated in a p-type well, there is no neutral p-type silicon region under the channel. Similarly, no neutral n-type region is under the channel of the p-MOSFET even though it is fabricated in an n-type well. Therefore

$$C_{dm} \to 0 \tag{10.18}$$

and

$$m \approx 1\,, \tag{10.19}$$

which results in near-ideal values of the subthreshold swing. At room temperature, the ideal subthreshold swing is

$$S \approx 2.3\frac{kT}{q} = 60 \text{ mV}\,. \tag{10.20}$$

Experimentally measured values of the subthreshold swing in FD SOI MOSFETs are typically 65 mV.

Example 10.5

Compare the standby dissipation for SOI CMOS (S = 60 mV) to that for bulk CMOS (S = 100 mV) at room temperature (300 K). Assume V_{DD} = 1 V, V_T = 0.1 V, and K = 1 mA/V² for both types of circuitry.

Solution. For the bulk CMOS, the subthreshold swing of 100 mV corresponds to m = 1.6. The standby dissipation per gate is

$$\begin{aligned} P_{subthreshold} &\approx V_{DD}K(1+m)\left(\frac{kT}{q}\right)^2 \exp\left(\frac{-V_T}{mkT/q}\right) \\ &= (1.0V)\left(10^{-3}\text{ A/V}^2\right)(1+1.6)\left(26\times10^{-3}\text{ V}\right)^2 \exp\left(\frac{-0.1\text{ V}}{42\text{ mV}}\right) \\ &= 1.6\times10^{-7}\text{ W}. \end{aligned}$$

For the FD SOI CMOS, the standby dissipation per gate is

$$\begin{aligned} P_{subthreshold} &\approx V_{DD}K(1+m)\left(\frac{kT}{q}\right)^2 \exp\left(\frac{-V_T}{mkT/q}\right) \\ &= (1.0V)\left(10^{-3}\text{ A/V}^2\right)\left(26\times10^{-3}\text{ V}\right)^2 \exp\left(\frac{-0.1\text{ V}}{26\text{ mV}}\right) \\ &= 1.4\times10^{-8}\text{ W}. \end{aligned}$$

Therefore, the FD SOI CMOS reduces the standby power compared to bulk CMOS by more than an order of magnitude, with all other things being equal.

10.9 Summary

Low-power CMOS integrated circuits are important for portable, battery-operated systems. In many CMOS circuits, the capacitance switching power is dominant, so the dissipation for a gate is given by

$$P \approx P_{switch} = \alpha f_{CLK} C_L V_{DD}^2 , \tag{10.21}$$

where V_{DD} is the supply voltage, C_L is the load capacitance, f_{CLK} is the clock frequency, and α is the switching activity factor. Aggressive scaling of the physical dimensions of the MOSFETs has led to significant reductions in the load capacitances and the dissipation per gate. Beyond that, the most effective way to reduce power is to reduce the supply voltage. However, supply voltage reduction must be accompanied by threshold voltage reduction in order to maintain acceptable propagation delays. This imposes a practical limit on the reduction of the supply voltage because a reduction of the device threshold voltages increases the standby power as a result of greater subthreshold conduction in the "cutoff" devices.

Adiabatic logic circuits conserve power by recycling electrical charge, whereas conventional CMOS gates use each electrical charge only once. In principle, this can be done by charging and discharging the load capacitance using current sources rather than V_{DD} and ground. During the discharge process, much of the energy that had been stored on the load capacitor can be returned to the supply for use by other logic circuits. Therefore, it is possible to reduce the switching dissipation to less than $fC_L V_{DD}^2$. In practical adiabatic circuits, the load capacitance can be charged and discharged stepwise, using several discrete supply voltages, and the switching dissipation can be reduced to 1/10 that of conventional CMOS.

The choice of supply voltage always involves a trade-off between speed and power, regardless of the choice of threshold voltage. In CMOS circuits operating with a single supply voltage, it is necessary to choose the supply voltage to meet the speed requirements in the most critical circuitry, often the I/O circuits. This results in wasted power in other parts of the integrated circuit.

Using two or more supply voltages in *multiple-voltage CMOS (MV CMOS)* alleviates this problem to a great extent because the output drivers can operate at a high supply voltage, for high speed, while the core of the integrated circuit can operate at a lower supply voltage for reduced dissipation. Nevertheless, the speed requirements vary significantly within the

core of the circuitry, from circuit to circuit and also over time. Therefore, the choice of any single supply voltage for the core of the circuitry will result in wasted power. *Dynamic voltage scaling (DVS)* allows the supply voltage to be adjusted dynamically for each block of circuitry. By making the supply voltage only adequate enough for the required throughput (speed), the power can be minimized.

Active body biasing avoids the trade-off between the standby power and speed performance, by allowing active adjustment of the threshold voltages in the circuit. This is done using the body bias effect, in which a bias voltage between the body and the source of the MOSFET adjusts its threshold voltage. Active circuits are biased to have small threshold voltages for improved speed, while inactive circuits are biased to have increased threshold voltages for reduced subthreshold conduction.

Multiple threshold CMOS (MT CMOS) circuits can be used to overcome the trade-off between speed and subthreshold conduction inherent in single threshold CMOS. Two or more distinct threshold voltages have been used in commercial microprocessors. The simplest multiple threshold scheme is dual threshold CMOS (DT CMOS), in which the logic circuits are disconnected from the supply rails during standby operation and thus can use low threshold voltages for optimum speed. High threshold MOSFETs are used to disconnect the logic circuits from the supply rails. These devices have low subthreshold leakage so that the standby power can be greatly reduced by this approach.

Silicon-on-insulator (SOI) is an important technology for low-power CMOS because SOI MOSFETs have superior subthreshold characteristics compared to bulk silicon MOSFETs. At the present time, SOI wafers are produced by the **s**eparation by the **im**plantation with **ox**ygen (SIMOX) and wafer bonding (WB) approaches. Both methods utilize the same wafers as bulk CMOS circuits, which are available in large diameter, with high quality, and at low cost. SOI MOSFETs may be partially depleted (PD) or fully depleted (FD). In FD SOI MOSFETs, the subthreshold slope is near the ideal value of 60 mV at room temperature. Combined with reduced process-induced variations in the threshold voltage, this allows implementation of SOI CMOS scaled down to V_{DD} = 0.4 V and V_T = 0.1 V. This reduction in the supply voltage decreases the capacitance switching power to one sixth of the value for bulk CMOS designed with V_{DD} = 1.0 V and V_T = 0.3 V. Further benefit derives from the reduction of the parasitic drain capacitances in SOI CMOS, which decreases the load capacitances (and therefore the switching dissipation) by about 20%.

LOW-POWER CMOS LOGIC QUICK REFERENCE

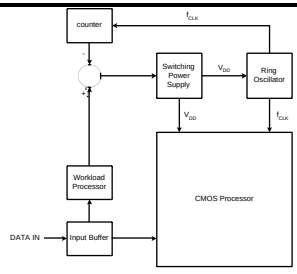
Dissipation in CMOS

$$P = \underbrace{P_{subthreshold} + P_{leakage}}_{P_{DC}} + \underbrace{P_{sc} + P_{switch}}_{P_{AC}}$$

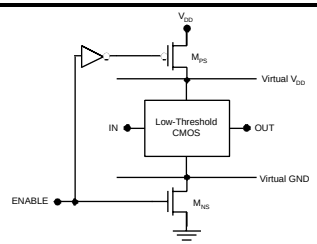
$$P_{switch} = \alpha f_{CLK} C_L V_{DD}^2 \qquad P_{subthreshold} \approx V_{DD} K (1+m) \left(\frac{kT}{q}\right)^2 \exp\left(\frac{-V_T}{mkT/q}\right)$$

Low power CMOS is important for portable applications. Device scaling and voltage reduction decrease the switching power. In LV CMOS, it is necessary to reduce the device thresholds to maintain acceptable speed. This increases the subthreshold conduction. Control of the subthreshold conduction in LV CMOS can be achieved using active body biasing, multiple threshold approaches, or fully depleted SOI. Because of the speed-power tradeoff in the choice of V_{DD}, multiple supply voltages and dynamic voltage scaling are commonly employed. Adiabatic logic gates recycle charges, reducing the switching power at a given supply voltage.

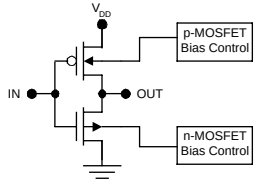
Dynamic Voltage Scaling (DVS)



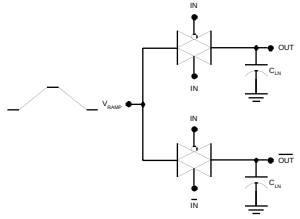
Multiple Threshold CMOS (MT CMOS)



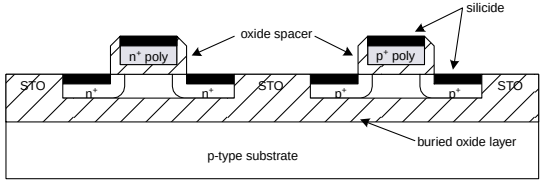
Active Body Biasing (ABB)



Adiabatic Logic Gates



Silicon-on-Insulator (SOI)



f = 10^{-15} p = 10^{-12} n = 10^{-9} μ = 10^{-6} m = 10^{-9} k = 10^{3} M = 10^{6} G = 10^{9}

Problems

P10.1. Suppose 1.0-V CMOS is fabricated using 0.25-μm technology and t_{OX} = 8 nm. Consider inverters with L_N = 0.25 μm, W_N = 1.0 μm, L_P = 0.25 μm, and W_P = 2.5 μm.

1. Calculate and plot the propagation delay (assuming 20 on-chip loads) vs. the threshold voltage (0.1 V ≤ V_T ≤ 0.6 V).
2. Calculate and plot the standby dissipation vs. the subthreshold voltage (0.1 V ≤ V_T ≤ 0.6 V). Assume the subthreshold conduction is dominant with S = 100 mV.

P10.2. Consider 0.25-μm CMOS technology with t_{OX} = 7 nm. Dual supply voltages are to be used. V_T = 0.3 V for all devices.

1. Choose V_{DDL} such that symmetric inverters with 20 on-chip loads can switch at 1 GHz (t_P < 500 ps).
2. Choose V_{DDH} such that output drivers with (W_P + W_N) < 50 μm can drive 15-pF loads at 250 MHz (t_P < 2 ns).

P10.3. Consider DVS implemented in 0.25-μm CMOS technology with t_{OX} = 7 nm. A ring oscillator is designed to mimic the critical path in the system using 13 stages of symmetric inverters with load capacitances at each stage equivalent to 15 on-chip loads.

1. Calculate and plot the ring oscillator frequency as a function of the supply voltage assuming VT = 0.3 V for all devices. 1 V ≤ V_{DD} ≤ 5 V.
2. Calculate and plot the switching power vs. the supply voltage for a symmetric inverter in the system. Assume the system clock frequency is derived from the ring oscillator, the switching activity is 0.2, and the on-chip fan-out is 15. For the symmetric inverter, W_N = 1 μm.

P10.4. Consider variable threshold CMOS fabricated in 0.25-μm technology with t_{ox} = 7 nm. The n-MOSFETs and the p-MOSFETs have 2 × 10^{16} cm^{-3} channel doping; ±3.3 V is available for active body biasing of the MOSFETs. The standby threshold voltages (with body bias) are to be ± 0.3 V. What is the minimum value for the absolute value of the nominal threshold voltages (as fabricated, without body bias)?

P10.5. Suppose dual threshold CMOS is implemented with thresholds of 0.1 and 0.3 V, and V_{DD} = 2.5 V. Consider symmetric inverters with L_N = 0.25 μm, W_N = 1.0 μm, L_P = 0.25 μm, W_P = 2.5 μm, and t_{OX} = 9 nm. Assume that symmetric inverters will be loaded with 15 fan-out gates and that each pair of switch transistors will be shared by 10 inverters.

1. Calculate the propagation delay for a low threshold inverter without power-switching transistors.
2. Determine the required widths for the power-switching transistors such that their addition increases the propagation delays by 10%.

P10.6. Show that, for an adiabatic logic gate that charges the load capacitance stepwise with n voltage steps, the capacitance switching dissipation is

$$P_{switch} = \frac{\alpha f_{CLK} V_{DD}^2 C_L}{n}.$$

References

1. Schrom, G. and Selberherr, S., Ultra-low-power CMOS technologies, *1996 Int. Semiconductor Conf.*, 237, 1996.
2. Sandararajan, V. and Parhi, K.K., Synthesis of low power CMOS VLSI circuits using dual supply voltages, *Proc. 36th Design Automation Conf.*, 72, 1999.
3. Qu, G., What is the limit of energy saving by dynamic voltage scaling? *IEEE/ACM 2001 Int. Conf. Computer Aided Design*, 560, 2001.
4. Nowka, K., Carpenter, G., Mac Donald, E., Ngo, H., Brock, B., Ishii, K., Nguyen, T., and Burns, J., A 0.9-V to 1.95-V dynamic voltage-scalable and frequency-scalable 32 b PowerPC processor, *Dig. Tech. Papers IEEE 2002 Int. Solid-State Circuits Conf.*, 340, 2002.
5. Burd, T.D., Pering, T.A., Stratakos, A.J., and Brodersen, R.W., A dynamic voltage scaled microprocessor system, *IEEE J. Solid-State Circuits*, 35, 1571, 2000.
6. Chung, E.-Y., Benini, L., and De Micheli, G., Contents provider-assisted dynamic voltage scaling for low energy multimedia applications, *Proc. 2002 Int. Symp. Low Power Electron. Design*, 42, 2002.
7. Kuroda, T. and Hamada, M., Low-power CMOS digital design with dual embedded adaptive power supplies, *IEEE J. Solid-State Circuits*, 35, 652, 2000.
8. Burd, T.D. and Brodersen, R.W., Design issues for dynamic voltage scaling, *Proc. 2000 Int. Symp. Low Power Electron. Design*, 9, 2000.
9. Sun, S.W. and Tsui, P.G.Y., Limitation of CMOS supply-voltage scaling by MOSFET threshold-voltage variation, *Proc. 1994 IEEE Custom Integrated Circuits Conf.*, 267, 1994.
10. Tschanz, J.W., Kao, J.T., Narendra, S.G., Nair, R., Antoniadis, D.A., Chandrakasan, A.P., and De, V., Adaptive body bias for reducing impacts of die-to-die and within-die parameter variations on microprocessor frequency and leakage, *IEEE J. Solid-State Circuits*, 37, 1396, 2002.
11. Kachi, T., Kaga, T., Wakahara, S., and Hisamoto, D., Variable threshold-voltage SOI CMOSFETs with implanted back-gate electrodes for power-managed low-power and high-speed sub-1-V ULSIs, *Dig. Tech. Papers 1996 Symp. VLSI Technol.*, 124, 1996.
12. Harada, M., Douseki, T., and Tsuchiya, T., Suppression of threshold voltage variation in MTCMOS/SIMOX circuit operating below 0.5 V, *Dig. Tech. Papers 1996 Symp. VLSI Technol.*, 96, 1996.

13. Inukai, T., Hiramoto, T., and Sakurai, T., Variable threshold voltage CMOS (VTCMOS) in series connected circuits, *2001 Int. Symp. Low Power Electron. Design*, 201, 2001.
14. Sakurai, T., Kawaguchi, H., and Kuroda, T., Low-power CMOS design through V_{TH} control and low-swing circuits, *Proc. 1997 Int. Symp. Low Power Electron. Design*, 1, 1997.
15. Wei, L., Chen, Z., Johnson, M., Roy, K., and De, V., Design and optimization of low voltage high performance dual threshold CMOS circuits, *Proc. 1998 Design Automation Conf.*, 489, 1998.
16. Tripathi, N., Bhosle, A., Samanta, D., and Pal, A., Optimal assignment of high threshold voltage for synthesizing dual threshold CMOS circuits, *14th Int. Conf. VLSI Design*, 227, 2001.
17. Adan, A.O., Naka, T., Kagisawa, A., and Shimizu, H., SOI as a mainstream IC technology, *Proc. 1998 SOI Conf.*, 9, 1998.
18. Assaderaghi, F. and Shahidi, G., SOI at IBM: current status of technology, modeling, design, and the outlook for the 0.1-µm generation, *Proc. 2000 IEEE Int. SOI Conf.*, 6, 2000.
19. Kawamura, S., Ultra-thin-film SOI technology and its application to next generation CMOS devices, *Proc. 1993 IEEE Int. SOI Conf.*, 6, 1993.
20. Hu, C., SOI and device scaling, *Proc. 1998 IEEE Int. SOI Conf.*, 1, 1998.
21. Aoki, T., Tomizawa, M., and Yoshii, A., Design considerations for thin-film SOI/CMOS device structures, *IEEE Trans. Electron. Devices*, 36, 1725, 1898.
22. Auberton–Herve, A.J., SOI: materials to systems, *1996 Int. Electron. Devices Meet.*, 3, 1996.
23. Mathew, S.K., Krishnamurthy, R.K., Anders, M.A., Rios, R., Mistry, K.R., and Soumyanath, K., Sub-500-ps 64-b ALUs in 0.18-µm SOI/bulk CMOS: design and scaling trends, *IEEE J. Solid-State Circuits*, 36, 1636, 2001.
24. Yoshino, A., Kumagai, K., Kurosawa, S., Itoh, H., and Okumura, K., Design methodology for low power, high-speed CMOS devices utilizing SOI technology, *Proc. 1993 IEEE Int. SOI Conf.*, 170, 1993.
25. Wei, L., Chen, Z., and Roy, K.; Design and optimization of double-gate SOI MOSFETs for low voltage low power circuits, *Proc. 1998 IEEE Int. SOI Conf.*, 69, 1998.
26. Fossum, J.G., Choi, J.-Y., and Sundaresan, R., SOI design for competitive CMOS VLSI, *IEEE Trans. Electron. Devices*, 37, 724, 1990.
27. Lee, J.-W., Kim, H.-K., Oh, J.-H., Yang, J.-W., Lee, W.-C., Kim, J.-S., Oh, M.-R., and Koh, Y.-H., A new SOI MOSFET for low power applications, *Proc. 1998 IEEE Int. SOI Conf.*, 65, 1998.
28. Shahidi, G., Ajmera, A., Assaderaghi, F., Bolam, R., Bryant, A., Coffey, M., Hovel, H., Lasky, J., Leobandung, E., Lo, H.-S., Maloney, M., Moy, D., Rausch, W., Sadana, D., Schepis, D., Sherony, M., Sleight, J.W., Wagner, L.F., Wu, K., Davari, B., and Chen, T., Mainstreaming of the SOI technology, *Proc. 1999 IEEE Int. SOI Conf.*, 1, 1999.
29. Pelella, M.M., Maszara, W., Sundararajan, S., Sinha, S., Wei, A., Ju, D., En, W., Krishnan, S., Chan, D., Chan, S., Yeh, P., Lee, M., Wu, D., Fuselier, M., vanBentum, R., Burbach, G., Lee, C., Hill, G., Greenlaw, D., Riccobenc, C., and Karlsson, O., Advantages and challenges of high performance CMOS on SOI, *Proc. 2001 IEEE Int. SOI Conf.*, 1, 2001.
30. Mathew, S., Krishnamurthy, R., Anders, M., Rios, R., Mistry, K., and Soumyanath, K., Sub-500 ps 64 b ALUs in 0.18-µm SOI/bulk CMOS: design and scaling trends, *Dig. Tech. Papers IEEE Int. Solid-State Circuits Conf.*, 318, 2001.

31. Colinge, J.P., Park, J.T., and Colinge, C.A., SOI devices for sub-0.1 μm gate lengths, *23rd Int. Conf. Microelectron.*, 109, 2002.
32. Alles, L., Dolan, P., Anc, M.J., Allen, P., Cordts, F., and Nakai, T., Analysis of ADVANTOX™ thin BOX SIMOX-SOI material, *Proc. 1997 IEEE Int. SOI Conf.*, 10, 1997.
33. Mizuno, T., Sugiyama, N., Kurobe, A., and Takagi, S., Advanced SOI p-MOSFETs with strained-Si channel on SiGe-on-insulator substrate fabricated by SIMOX technology, *IEEE Trans. Electron. Devices*, 48, 1612, 2001.
34. Liu, S.T., Jenkins, W., Hughes, H., and Auberton–Herve, A.J., Radiation properties of UNIBOND™ with 200 nm buried oxide [SOI wafers], *Proc. 1998 IEEE Int. SOI Conf.*, 93, 1998.
35. Maleville, C., Barge, T., Ghyselen, B., Auberton, A.J., Moriceau, H., and Cartier, A.M., Multiple SOI layers by multiple Smart-Cut® transfers, *Proc. 2000 IEEE Int. SOI Conf.*, 134, 2000.
36. Neuner, J.W., Ledger, A.M., Schilb, S.K., and Mathur, D.P., Improved uniformity in bonded SOI wafers with active layers from 1 to 30 μm at high throughputs, *Proc. 1998 IEEE Int. SOI Conf.*, 169, 1998.
37. Ito, M., Yamagata, K., Miyabayashi, H., and Yonehara, T., Scalability potential in ELTRAN® SOI-epi wafer, *Proc. 2000 IEEE Int. SOI Conf.*, 10, 2000.
38. Yeh, W.K., Huang, C., Chen, T.F., Hsu, S.M., Liu, J., Lin, C.H., and Liou, F.T., High performance 0.1 μm partially depleted SOI CMOSFET, *Proc. 2000 IEEE Int. SOI Conf.*, 68, 2000.
39. Assaderaghi, F., Shahidi, G., Fung, S., Sherony, M., Wagner, L., Sleight, J., Lo, S.H., Wu, K., and Chen, T.-C., Partially depleted silicon-on-insulator (SOI): a device design/modeling and circuit perspective, *Proc. 12th Int. Conf. Microelectron.*, 201, 2000.
40. Pelella, M.M., Fossum, J.G., and Krishnan, S., Control of off-state current in scaled PD/SOI CMOS digital circuits, *Proc. 1998 IEEE Int. SOI Conf.*, 147, 1998.
41. Numata, T., Noguchi, M., Oowaki, Y., and Takagi, S., Back gate engineering for suppression of threshold voltage fluctuation in fully-depleted SOI MOSFETs, *Proc. 2000 IEEE Int. SOI Conf.*, 78, 2000.
42. Brady, F.T. and Haddad, N.F., Manufacturability considerations for fully depleted SOI, *Proc. 1993 IEEE Int. SOI Conf.*, 130, 1993.
43. Yeh, P.C. and Fossum, J.G., Viable deep-submicron FD/SOI CMOS design for low-voltage applications, *Proc. 1994 IEEE Int. SOI Conf.*, 23, 1994.
44. Zhang, R., Roy, K., and Janes, D.B., Double-gate fully depleted SOI transistors for low-power high-performance nano-scale circuit design, *2001 Int. Symp. Low Power Electron. Design*, 213, 2001.
45. Kyriakis-Bitzaros, E.D. and Nikolaidis, S.S., Design of low power CMOS drivers based on charge recycling, *Proc. 1997 IEEE Int. Symp. Circuits Syst.*, 3, 1924, 1997.
46. Liu, F. and Lau, K.T., Pass-transistor adiabatic logic with NMOS pull-down configuration, *Electron. Lett.*, 34, 739, 1998.
47. Lim, J., Kim, D.-G., and Chae, S.-I., nMOS reversible energy recovery logic for ultra-low-energy applications, *IEEE J. Solid-State Circuits*, 35, 865, 2000.

11

BiCMOS Logic

11.1 Introduction

CMOS is preferred for most applications because of its low standby dissipation, high packing density, rail-to-rail voltage swing, and excellent speed. However, because bipolar circuitry still outperforms CMOS in terms of off-chip data rates, bipolar-CMOS (BiCMOS) logic was developed in an attempt to achieve the advantages of both types of logic gates. At the present time, high-performance BiCMOS circuits have been developed based on Si and SiGe technology and are widely used.[1–15]

A number of different versions of BiCMOS logic gates are available[9–15]; however, they all share the salient features of the inverter shown in Figure 11.1. The two important design features of this circuit are 1) CMOS-type logic circuitry and 2) a totem-pole output formed using two npn bipolar transistors. The CMOS logic circuitry provides low standby power dissipation while the bipolar transistors at the output provide superior performance with highly capacitive loads. In fact, the packing density of BiCMOS can be excellent because the relatively large bipolar transistors are only needed to drive the output connections; thus, BiCMOS integrated circuits are really just CMOS on the inside.

11.2 Voltage Transfer Characteristic

Consider the voltage transfer characteristic for a BiCMOS inverter. With logic zero (0 V) applied at the input as shown in Figure 11.2, M_{N1} and M_{N3} are cut off while M_{P1} is linear. As a result, the voltage at the base of Q_P is equal to V_{DD} and M_{N2} is linear, Q_O is cut off, and Q_P is forward active. From the diagram it is clear that the output high voltage is 0.7 V less than the supply voltage:

$$V_{OH} = V_{DD} - V_{BEA}\,. \tag{11.1}$$

Next consider the case of logic one (V_{DD}) applied at the input as shown in Figure 11.3. For this situation, M_{N1} and M_{N3} are linear while M_{P1} is cut off.

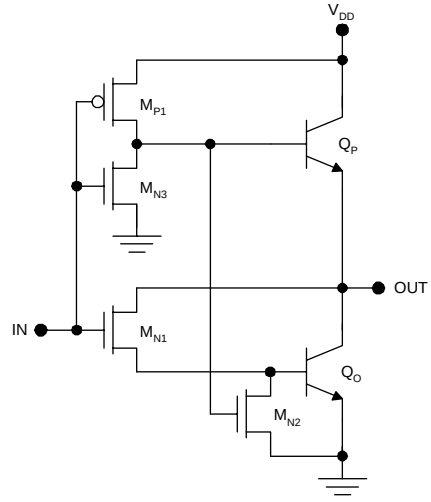


FIGURE 11.1
BiCMOS inverter.

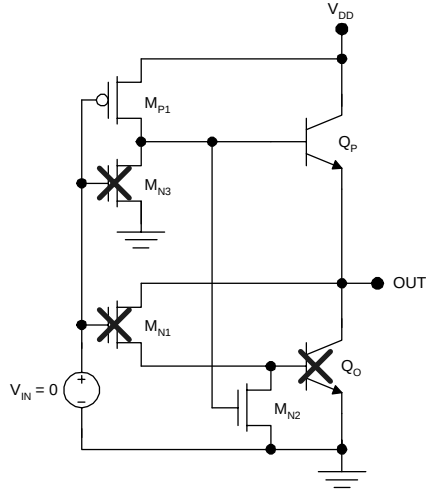


FIGURE 11.2
BiCMOS inverter with logic zero applied at the input. The devices marked with "X" are cut off.

Q_P will be cut off because the voltage at the base is zero; M_{N2} will also be cut off, so Q_O will be forward active. From the diagram it is clear that the output low voltage is about 0.7 V:

$$V_{OL} = V_{BEA} . \tag{11.2}$$

Therefore, the logic swing of conventional BiCMOS circuits is about 1.4 V less than the supply voltage:

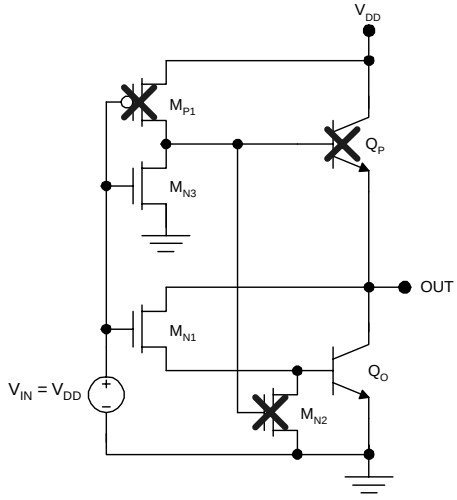


FIGURE 11.3
BiCMOS inverter with logic one applied at the input. The devices marked with "X" are cut off.

$$V_{OH} - V_{OL} = V_{DD} - 2V_{BEA}\,. \tag{11.3}$$

This fact has important implications for the future of BiCMOS. The relentless drive to lower supply voltages makes this 1.4-V degradation increasingly important. For example, with a 1.8-V supply, the logic swing of BiCMOS is a mere 0.4 V. As a consequence, conventional BiCMOS circuits are undesirable for $V_{DD} < 2.5$ V.

11.3 Propagation Delays

The motivation for using BiCMOS over CMOS is improved performance with highly capacitive loads. It is therefore important to compare the propagation delays of BiCMOS and CMOS. Consider the CMOS and BiCMOS inverters shown in Figure 11.4. For the sake of a fair comparison, it is necessary to assume that the MOSFETs in the two circuits are fabricated with identical K and V_T values. For the CMOS inverter, the propagation delay is

$$t_P(CMOS) \approx \frac{C_L}{K}\left[\frac{2V_T}{\left(V_{DD}-V_T\right)^2} + \frac{2}{\left(V_{DD}-V_T\right)}\ln\left(\frac{V_{DD}-V_T}{V_{DD}/2}\right)\right]. \tag{11.4}$$

Calculation of the propagation delay for the BiCMOS gate is rather complex and best done by SPICE. However, for the purpose of hand calculations,

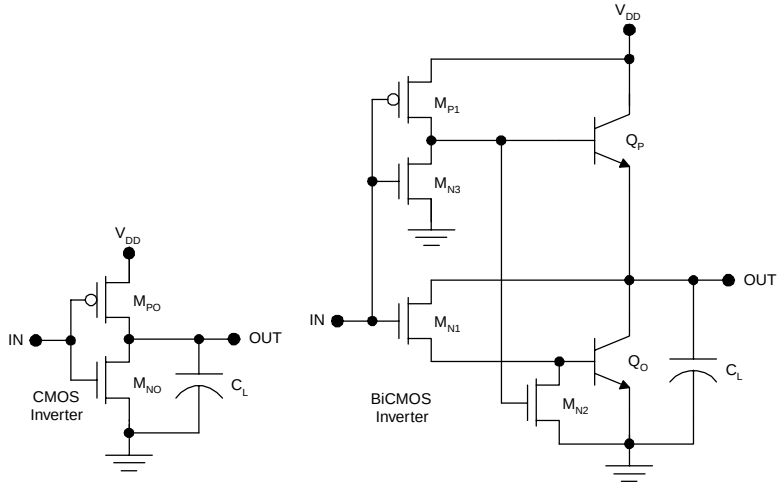


FIGURE 11.4
CMOS and BiCMOS inverters with lumped capacitive loads.

the BiCMOS propagation delay can be estimated as the sum of the two dominant components as follows. The first component is the propagation delay of the M_{P1}/M_{N3} CMOS inverter loaded by the Q_P emitter follower. The load capacitance seen by the M_{P1}/M_{N3} inverter is approximately equal to the base-collector capacitance of Q_P. The second component is associated with the RC time constant at the output. The output impedance of the emitter follower is approximately equal to $1/g_m$, where g_m is the transconductance for Q_P ($1/g_m$ is on the order of a few ohms). Adding these two components, the BiCMOS propagation delay is approximately

$$t_P(BiCMOS) \approx \frac{C_{BC}}{K}\left[\frac{2V_T}{(V_{DD}-V_T)^2}+\frac{2}{(V_{DD}-V_T)}\ln\left(\frac{V_{DD}-V_T}{V_{DD}/2}\right)\right]+\ln(2)\frac{C_L}{g_m}. \quad (11.5)$$

Typically, the second term is negligible except for large values of C_L. In such cases,

$$\frac{t_P(CMOS)}{t_P(BiCMOS)} \approx \frac{C_L}{C_{BC}}. \quad (11.6)$$

It is therefore beneficial to use BiCMOS if $C_L > C_{BC}$, which is typically the case for off-chip loads.* However, the use of CMOS is indicated for cases in which $C_L < C_{BC}$, a situation that exists for on-chip loads.**

* Off-chip loads are typically 10 to 100 pF, whereas a typical value for C_{BC} is less than 1 pF.
** On-chip loads are typically less than 100 fF, except for lines with high fan-out and long runs (such as clock lines).

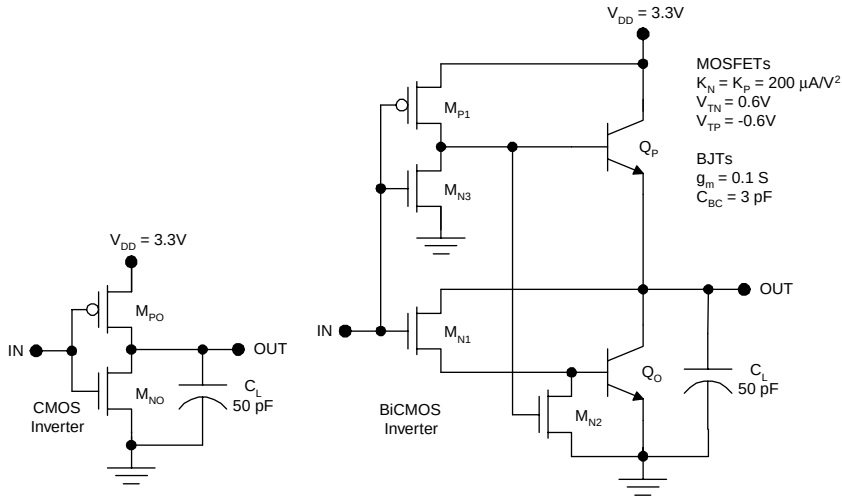


FIGURE 11.5
Comparison of propagation delays for CMOS and BiCMOS with 50-pF off-chip loads.

Example 11.1

Compare the propagation delays of the CMOS and BiCMOS gates with 50-pF loads as illustrated in Figure 11.5.

Solution. For the CMOS gate,

$$t_p(CMOS) \approx \frac{50\times10^{-12}\,\text{F}}{0.4\times10^{-3}\,\text{A/V}^2}\left[\frac{1.2\,\text{V}}{(2.7\,\text{V})^2}+\frac{2}{2.7\,\text{V}}\ln\left(\frac{2.7\,\text{V}}{1.65\,\text{V}}\right)\right]=66\,\text{ns}.$$

For the BiCMOS gate,

$$t_p(BiCMOS) \approx \frac{3\times10^{-12}\,\text{F}}{0.4\times10^{-3}\,\text{A/V}^2}\left[\frac{1.2\,\text{V}}{(2.7\,\text{V})^2}+\frac{2}{2.7\,\text{V}}\ln\left(\frac{2.7\,\text{V}}{1.65\,\text{V}}\right)\right]$$

$$+\ln(2)\frac{50\times10^{-12}\,\text{F}}{0.1S}=4.3\,\text{ns}$$

The BiCMOS gate is 15 times faster than the CMOS gate for the case of 50-pF off-chip loads.

Example 11.2

Compare the propagation delays of the CMOS and BiCMOS gates shown in Figure 11.6 for the case of 10 on-chip CMOS loads.

Solution. For the n-MOSFETs and p-MOSFETs, the process transconductance parameters are

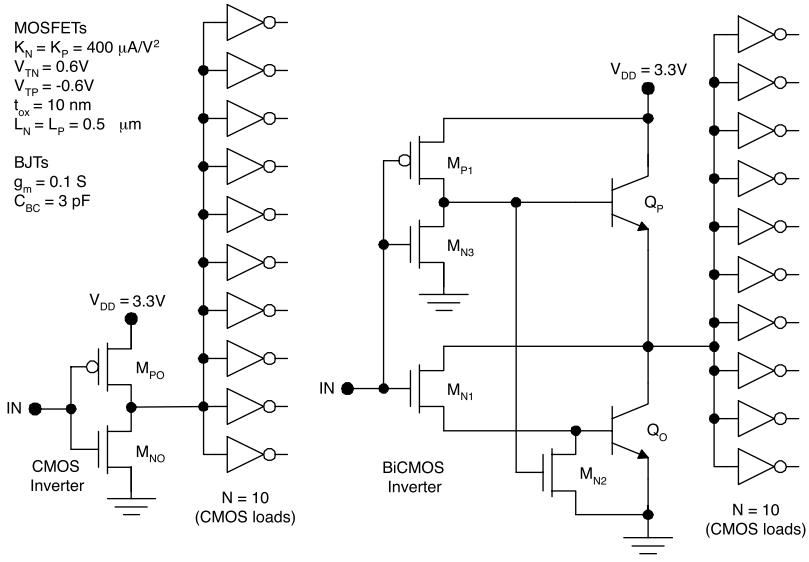


FIGURE 11.6
CMOS and BiCMOS gates driving 10 on-chip CMOS load gates.

$$k'_N = \frac{\varepsilon_{ox}\mu_N}{t_{ox}} = \frac{(3.9)(8.85\times10^{-14}\ \text{F/cm})(580\ \text{cm}^2/\text{Vs})}{10\times10^{-7}\ \text{cm}} = 0.2\ \text{mA/V}^2$$

and

$$k'_P = \frac{\varepsilon_{ox}\mu_P}{t_{ox}} = \frac{(3.9)(8.85\times10^{-14}\ \text{F/cm})(230\ \text{cm}^2/\text{Vs})}{10\times10^{-7}\ \text{cm}} = 0.079\ \text{mA/V}^2\ ,$$

respectively. Both devices have half-micron gate lengths and the device transconductance parameters are 0.4 mA/V^2. Therefore, the device widths are

$$W_N = L_N \frac{K_N}{k'_N} = 0.5\ \mu\text{m}\,\frac{0.4\ \text{mA/V}^2}{0.2\ \text{mA/V}^2} = 1.0\ \mu\text{m}$$

and

$$W_P = L_P \frac{K_P}{k'_P} = 0.5\ \mu\text{m}\,\frac{0.4\ \text{mA/V}^2}{0.079\ \text{mA/V}^2} = 2.5\ \mu\text{m}\,.$$

The input capacitance for each CMOS load gate is

$$C_{IN} = C_{OXN} + C_{OXP} = \frac{\varepsilon_{ox} W_N L_N}{t_{ox}} + \frac{\varepsilon_{ox} W_P L_P}{t_{ox}} = 1.73fF + 4.3fF = 6.03fF .$$

With 10 such CMOS loads, the load capacitance is

$$C_L = 10C_{IN} = 60fF .$$

For the CMOS driving gate,

$$t_p(CMOS) \approx \frac{60\times10^{-15} F}{0.4\times10^{-3}\,\mathrm{A/V^2}}\left[\frac{1.2\text{ V}}{(2.7\text{ V})^2} + \frac{2}{2.7\text{ V}}\ln\left(\frac{2.7\text{ V}}{1.65\text{ V}}\right)\right] = 79\text{ ps} .$$

and for the BiCMOS driving gate,

$$t_p(BiCMOS) \approx \frac{3\times10^{-12}\,\mathrm{F}}{0.4\times10^{-3}\,\mathrm{A/V^2}}\left[\frac{1.2\text{ V}}{(2.7\text{ V})^2} + \frac{2}{2.7\text{ V}}\ln\left(\frac{2.7\text{ V}}{1.65\text{ V}}\right)\right]$$

$$+\ln(2)\frac{60\times10^{-15}\,\mathrm{F}}{0.1S} = 4.0\text{ ns}$$

For the case of 10 on-chip CMOS load gates, the CMOS is actually 50 times faster than the BiCMOS. This shows that BiCMOS is useful only for driving highly capacitive loads, such as off-chip loads or on-chip loads with high fan-out.

11.4 Rail-to-Rail BiCMOS

A key limitation of BiCMOS is the reduced logic swing compared to CMOS. Because of this numerous attempts have been made to achieve rail-to-rail BiCMOS. The simplest way to achieve this operation of BiCMOS is to use a parallel CMOS output driver as shown in Figure 11.7. Although this gate achieves rail-to-rail operation, the dynamic rail-to-rail performance becomes limited by the MOSFETs. In other words, Q_P provides no benefit to the dynamic performance for $V_{DD} - V_{BEA} \le V_{OUT} \le V_{DD}$ because the bipolar transistor is cut off for this range of V_{OUT}. Similarly, Q_O provides no benefit to the dynamic performance for $0 \le V_{OUT} \le V_{BEA}$ because that bipolar transistor is cut off for this range of V_{OUT}. For example, with a 1.8-V supply voltage, the bipolar transistors would conduct only for the range 0.7 V ≤ V_{OUT} ≤ 1.1 V. As a consequence, this version of rail-to-rail BiCMOS offers no real advantage over CMOS for supply voltages less than about 2.5 V.

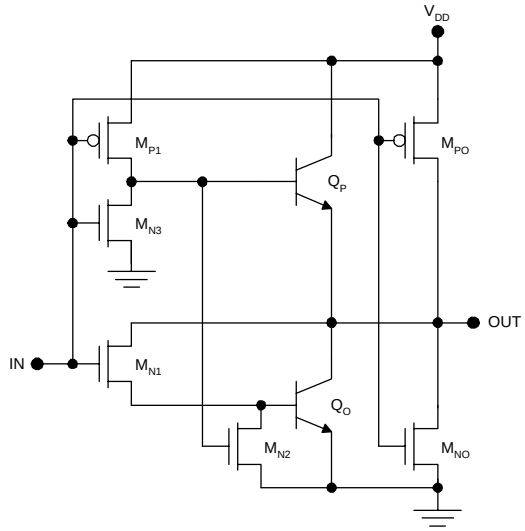


FIGURE 11.7
Rail-to-rail BiCMOS using a paralleled CMOS circuit.

Another strategy for achieving rail-to-rail BiCMOS is to use passive shunts as shown in Figure 11.8. Here, two passive (resistive) shunts are placed across the base–emitter junctions of the bipolar transistors. The shunt resistor R_1 allows the output voltage to swing all the way to V_{DD} while Q_P is cut off. Dynamically, this means that the load capacitance must charge through this resistance for $V_{OUT} > V_{DD} - V_{BEA}$. Similarly, R_2 allows the output to swing all the way to zero, but with a relatively poor slew rate.

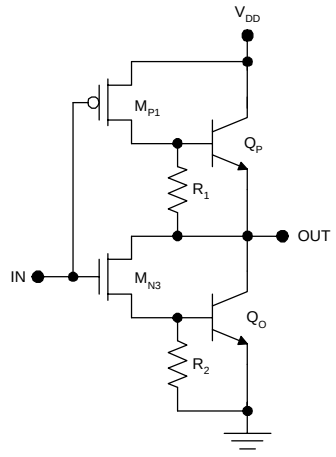


FIGURE 11.8
Rail-to-rail BiCMOS inverter with passive shunts.

Other versions of rail-to-rail BiCMOS exist, but they all suffer from this same basic limitation. It appears that shunting the bipolar transistors is not the answer! After all, bipolar transistors cannot boost the dynamic performance of the gate unless they are conducting. The idea behind shunting is that it allows the bipolar transistors to turn off, while shunt elements allow the base-emitter voltage to approach zero. All of this seems to indicate a fundamental limitation of BiCMOS that precludes its use in low-voltage circuits. However, BiCMOS has been used in dual voltage applications. The idea behind dual-voltage BiCMOS is to use a low voltage supply for the CMOS core, thus minimizing the power dissipation, while using a higher voltage supply for the BiCMOS output drivers, for greater logic swing and improved off-chip bit rates. This practical solution has been implemented in commercial products such as microprocessors.

11.5 Logic Design

In BiCMOS circuitry, the NAND function is formed by paralleling the p-channel MOSFETs and forming series combinations of n-channel MOSFETs. For example, the two-input NAND (NAND2) gate is shown in Figure 11.9. If either input goes low, the associated M_{P1} will conduct while the associated M_{N1} and M_{N3} will be cut off. This will cause Q_P to conduct and Q_O to be cut off, so the output will go high. On the other hand, if both inputs are brought to logic one, the output will go low.

The three-input BiCMOS NAND gate is realized by placing three p-MOSFETs in parallel and forming series combinations of the n-channel MOSFETs as shown in Figure 11.10. The M-input BiCMOS NAND gate requires M p-MOSFETs in parallel and two series combinations of M n-MOSFETs, as well as two bipolar transistors and the discharge MOSFET M_{N2}. Therefore, the M-input BiCMOS NAND gate requires (3M + 1) MOSFETs and two bipolar transistors. In order to maintain the same performance as that for the inverter, the series n-channel MOSFETs must be scaled by a factor of M (i.e., their *K* values must be increased by a factor of M). This preserves the output low, on-state resistance between V_{OUT} and ground, as well as the t_{PHL} performance. However, the paralleled p-MOSFETs need not be scaled compared to the inverter.

The two-input BiCMOS NOR gate is implemented by placing the p-MOSFETs in series and forming parallel combinations of n-MOSFETs. The two-input version is shown in Figure 11.11.

The M-input BiCMOS NOR gate requires (3M + 1) MOSFETs and two bipolar transistors, just as the M-input NAND gate does. However, *for the NOR gate, the p-MOSFETs must be scaled up by a factor of M*. This means that the K values of the p-MOSFETs must be increased by a factor of M compared to the inverter with similar performance. In practice, this scaling is achieved by increasing the gate widths while leaving the gate lengths unchanged. The

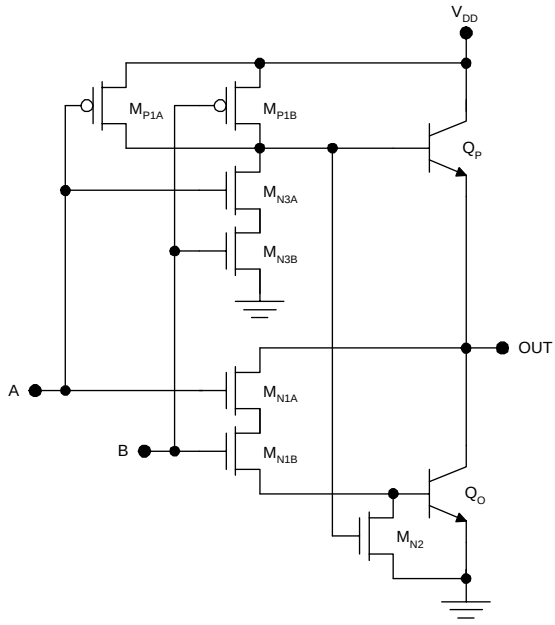


FIGURE 11.9
BiCMOS NAND2 gate.

n-MOSFETs need not be scaled in the BiCMOS NOR gate. As with CMOS, NAND implementations are preferred over NOR implementations due to the better layout efficiency of the NAND gate.

More complex AND–OR–INVERT functions may be implemented in BiCMOS by forming more complex pull-up and pull-down networks. The design methodology and the scaling rules are the same as for the case of CMOS and will not be elaborated here.

11.6 PSPICE Simulations

Simulations were performed using PSPICE 9.1, which can be downloaded free.[16] The BJT model parameters used in all simulations are provided in Table 11.1. The MOSFET model parameters used in all simulations are provided in Table 11.2 and Table 11.3.

11.6.1 Voltage Transfer Characteristic

The voltage transfer characteristic was simulated for the BiCMOS inverter of Figure 11.12 with V_{DD} = 3.3 V. Two 1-MΩ load resistors were connected

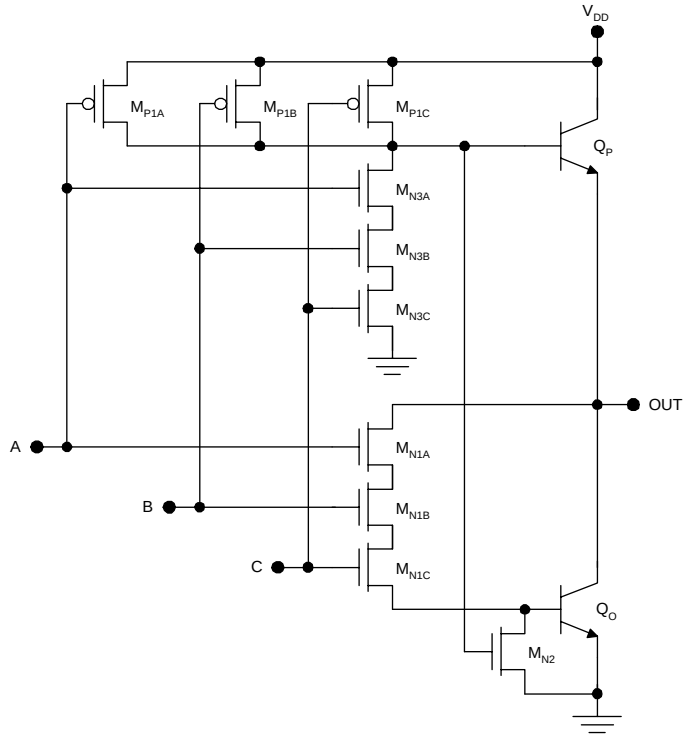


FIGURE 11.10
BiCMOS NAND3 gate.

to guarantee minimal DC loading; this results in realistic base-emitter voltage drops in the bipolar transistors. The resulting voltage transfer characteristic is shown in Figure 11.13; the observed logic swing is 2.22 V, or 1.08 V less than the supply voltage.

11.6.2 Propagation Delays

The propagation delays for a BiCMOS inverter with a 15-pF load were determined using the circuit of Figure 11.14. The results of the transient simulation appear in Figure 11.15. The propagation delays can be determined using the 50% points on the input and output waveforms and are t_{PHL} = 1.31 ns and t_{PLH} = 1.67 ns. Additional transient simulations were done to determine the propagation delays as a function of the load capacitance and compare these results to those for CMOS. The results are shown in Figure 11.16. The BiCMOS circuitry outperforms CMOS unless the load capacitance is so small that it is smaller than or comparable to the parasitic capacitances in the bipolar transistors.

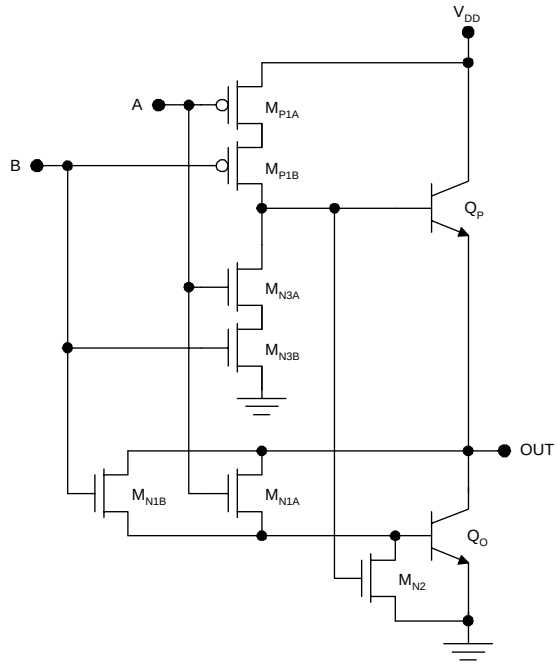


FIGURE 11.11
BiCMOS NOR2 gate.

TABLE 11.1

BJT SPICE Parameters for PSPICE Simulations

Parameter	Value	Units
IS	2.0f	A
BF	70	—
NF	1	—
BR	0.5	—
NR	1	—
CJE	0.3p	F
VJE	0.8	V
MJE	0.333	—
TF	0.2n	s
CJC	0.15p	F
VJC	0.75	V
MJC	0.5	—
TR	10n	s

TABLE 11.2

n-MOSFET SPICE Parameters for PSPICE Simulations

Parameter	Value	Units
VTO	0.5	V
KP	0.2m	A/V^2
LAMBDA	0.05	—
CGSO	1.15n	F/m
CGDO	0.58n	F/m
VMAX	100k	m/s

TABLE 11.3

p-MOSFET SPICE Parameters for PSPICE Simulations

Parameter	Value	Units
VTO	–0.5	V
KP	0.2m	A/V^2
LAMBDA	0.05	—
CGSO	1.15n	F/m
CGDO	0.58n	F/m
VMAX	80k	m/s

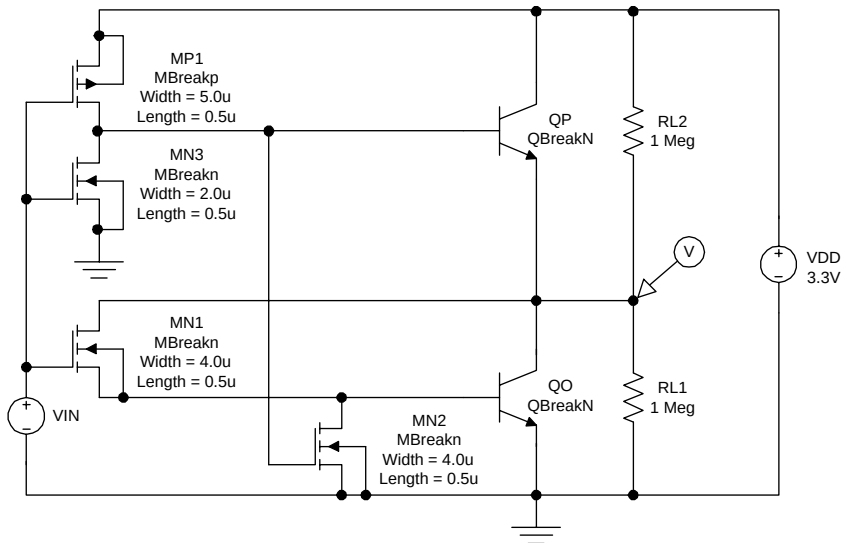


FIGURE 11.12
BiCMOS circuit used for simulation of the VTC.

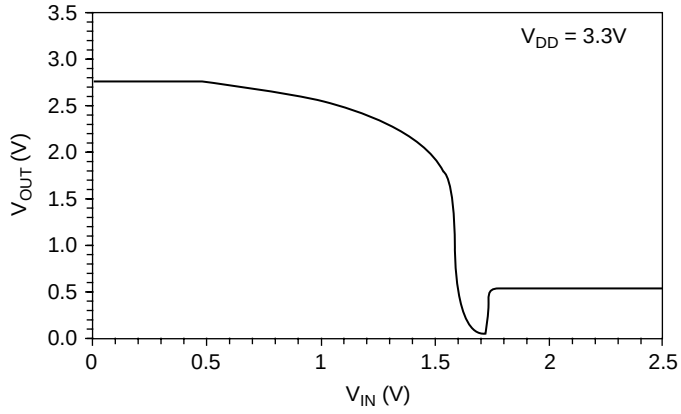


FIGURE 11.13
Simulated VTC for the BiCMOS circuit.

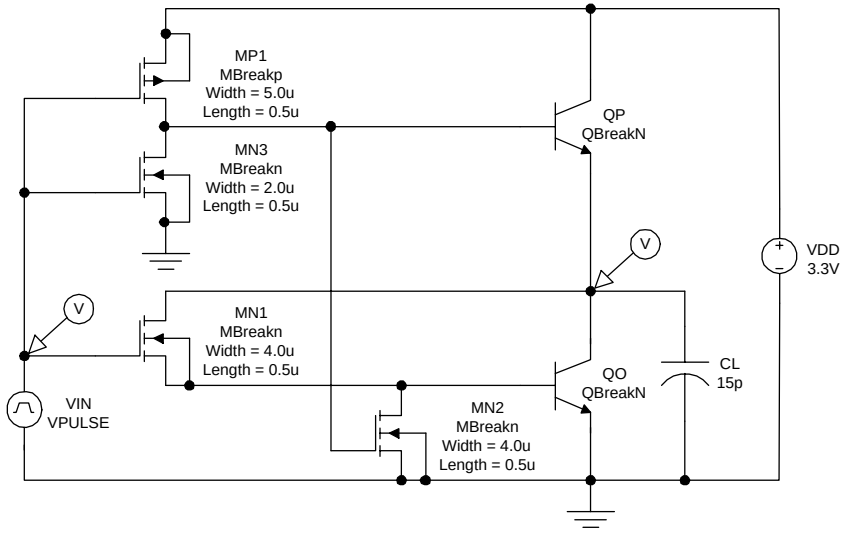


FIGURE 11.14
BiCMOS circuit used for simulation of the propagation delays.

11.7 Summary

BiCMOS circuits combine MOSFETs and bipolar transistors in order to achieve the low power dissipation of CMOS and the high off-chip data rates of bipolar logic gates. BiCMOS circuits use pure CMOS gates internally but the output drivers have bipolar totem pole outputs. A bipolar driver stage

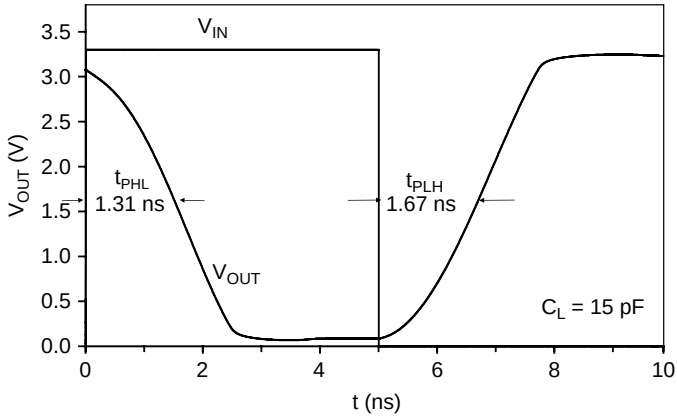


FIGURE 11.15
Simulated BiCMOS transient response.

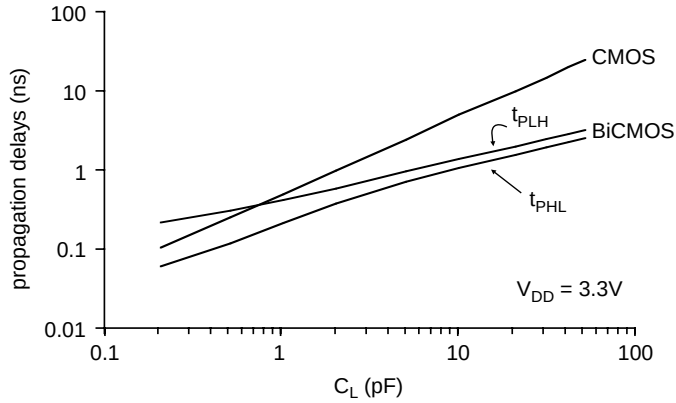


FIGURE 11.16
Comparison of BiCMOS and CMOS propagation delays.

has a very low output impedance, resulting in a low RC product associated with the load.

An important disadvantage of BiCMOS is the reduced logic swing compared to CMOS, which is becoming a critical problem because of the need to reduce supply voltages and power densities. *Rail-to-rail BiCMOS circuits* have not solved this problem adequately because *these rail-to-rail circuits achieve the improved logic swing without addressing the dynamic performance*.

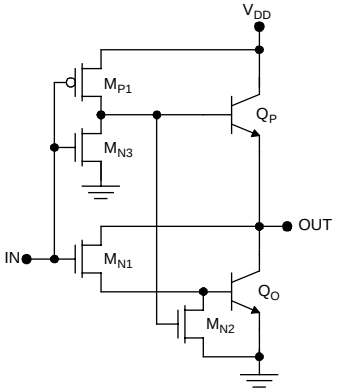
The speed advantage of BiCMOS is significant for the case of off-chip loads and moderate supply voltages; however, BiCMOS circuits are actually slower for the cases of on-chip loads and low supply voltages. Therefore, BiCMOS chips are typically implemented using CMOS to realize the internal logic circuitry and the BiCMOS circuits are reserved for use in high-voltage output drivers and a few other special-purpose circuits with high load capacitances.

Advances in fabrication have continually shrunk the dimensions of transistors, boosting the performance of MOSFETs relative to bipolar transistors. This has made buffered CMOS drivers an attractive alternative to BiCMOS. A further advantage of buffered CMOS is that it avoids the extra processing steps necessary to achieve bipolar transistors on the same chip with MOSFETs.

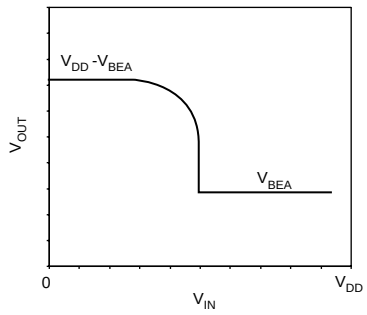
The future of BiCMOS is unclear. However, ever-decreasing supply voltages reveal that MOSFETs have a fundamental advantage over bipolar devices. Whereas the threshold voltage for a silicon-on-insulator n-MOSFET can be reduced to about 0.1 V (limited by the subthreshold turn-off characteristic), the turn-on voltage for a silicon p–n junction is about 0.7 V. MOSFETs therefore seem poised to win the battle for low-voltage, low-power integrated circuits.

BiCMOS LOGIC QUICK REFERENCES

Inverter Circuit



DC Voltage Transfer Characteristic



BiCMOS logic circuits are designed to combine the high packing density and low power of CMOS with the high off-chip data rates of bipolar logic. The primary disadvantage compared to CMOS is the reduced voltage swing.

DC Voltage Transfer Characteristic

$$V_{OL} = V_{BEA} \qquad V_{OH} = V_{DD} - V_{BEA}$$

Propagation Delays

$$t_P(BiCMOS) \approx \frac{C_{BC}}{K}\left[\frac{2V_T}{(V_{DD}-V_T)^2} + \frac{2}{(V_{DD}-V_T)}\ln\left(\frac{V_{DD}-V_T}{V_{DD}/2}\right)\right] + \ln(10)\frac{C_L}{g_m}$$

Comparison with CMOS

$$\frac{[V_{OH}-V_{OL}]_{BiCMOS}}{[V_{OH}-V_{OL}]_{CMOS}} = \frac{V_{DD}-2V_{BEA}}{V_{DD}} \qquad \frac{t_P(CMOS)}{t_P(BiCMOS)} \approx \frac{C_L}{C_{BC}}$$

Design Rules

$$V_{BEA} = 0.7V \qquad \mu_n = 580cm^2V^{-1}s^{-1} \qquad \mu_p = 230cm^2V^{-1}s^{-1} \qquad \varepsilon_{OX} = 3.9\varepsilon_0$$

f = 10^{-15} p = 10^{-12} n = 10^{-9} μ = 10^{-6} m = 10^{-3} k = 10^{3} M = 10^{6} G = 10^{9}

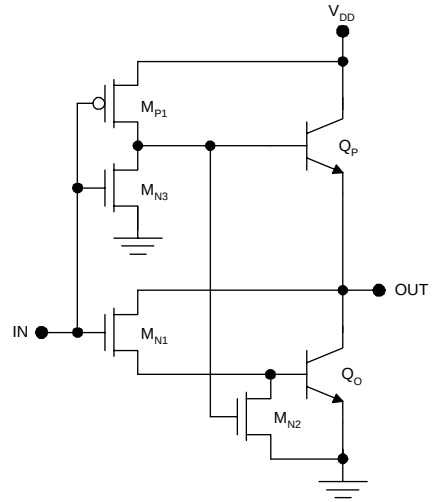


FIGURE 11.17
BiCMOS inverter (L11.1).

Laboratory Exercises

L11.1. Obtain commercially available bipolar transistors and enhancement type n-channel and p-channel MOSFETs for the construction of a BiCMOS inverter as shown in Figure 11.17. From the manufacturer's Web pages, obtain the data sheets for all three types of transistors.

1. Determine and plot the *unloaded* voltage transfer characteristic, using hand calculations, for the case of V_{DD} = 5 V.
2. Using SPICE, determine and plot the *unloaded* voltage transfer characteristics with V_{DD} as a parameter. Consider cases in which V_{DD} = 5 V, 3.3 V, 2.5 V, 1.8 V, and 1.5 V.
3. Build the inverter and measure the unloaded voltage transfer characteristics using the x–y feature of an oscilloscope or virtual instrument and a low-frequency input signal (1 kHz). Consider cases in which V_{DD} = 5 V, 3.3 V, 2.5 V, 1.8 V, and 1.5 V.
4. Plot the logic swing vs. the supply voltage from the SPICE and experimental results. What can be concluded from these results?

L11.2. Obtain commercially available bipolar transistors and enhancement type n-channel and p-channel MOSFETs for the construction of

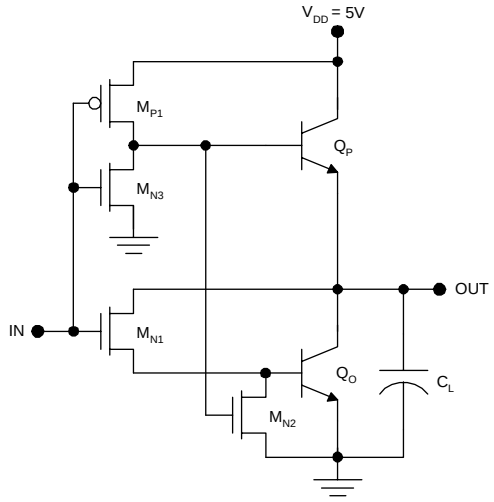


FIGURE 11.18
BiCMOS inverter (L11.2).

BiCMOS inverter as shown in Figure 11.18. From the manufacturer's Web pages, obtain the data sheets for all three types of transistors.

1. Using SPICE, estimate the load capacitance, $C_{L,1000}$, for which the average propagation delay will be 1000 ns.
2. Using SPICE, determine and plot the average propagation delay as a function of the load capacitance for $C_{L,1000}/10 \le C_L \le 10C_{L,1000}$.
3. Build a three-stage ring oscillator using these inverter circuits. Measure the average propagation delay as a function of the load capacitance, for $C_{L,1000}/10 \le C_L \le 10C_{L,1000}$.
4. Plot the SPICE and experimental results on the same graph. Are the characteristic linear? Do the extrapolated characteristics pass through the origin? Explain.

L11.3. Obtain commercially available bipolar transistors and enhancement type n-channel and p-channel MOSFETs for the construction of BiCMOS and CMOS inverters for the construction of three-stage ring oscillators as shown in Figure 11.19. From the manufacturer's Web pages, obtain the data sheets for all three types of transistors.

1. Using SPICE, estimate the load capacitance, $C_{L,100}$, for which the average propagation delay will be 100 ns for the BiCMOS gate with a 5-V supply.
2. Build the BiCMOS ring oscillator using this value of capacitance. Measure and plot the average propagation delay vs. the supply voltage for the range 2 V $\le V_{DD} \le$ 6 V.

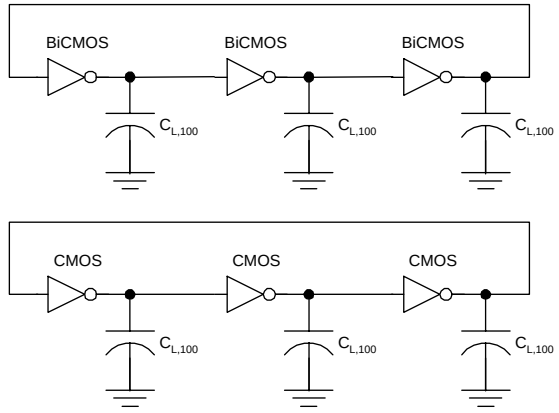


FIGURE 11.19
BiCMOS and CMOS three-stage ring oscillators (L11.3).

3. Using the CMOS ring oscillator, measure and plot the average propagation delay vs. the supply voltage for the range 2 V ≤ V_{DD} ≤ 6 V. Use the same load capacitance used for the BiCMOS measurements.
4. Compare and contrast the BiCMOS and CMOS results.

L11.4. Obtain commercially available enhancement type n-channel and p-channel MOSFETs for construction of the rail-to-rail BiCMOS inverter shown in Figure 11.20.

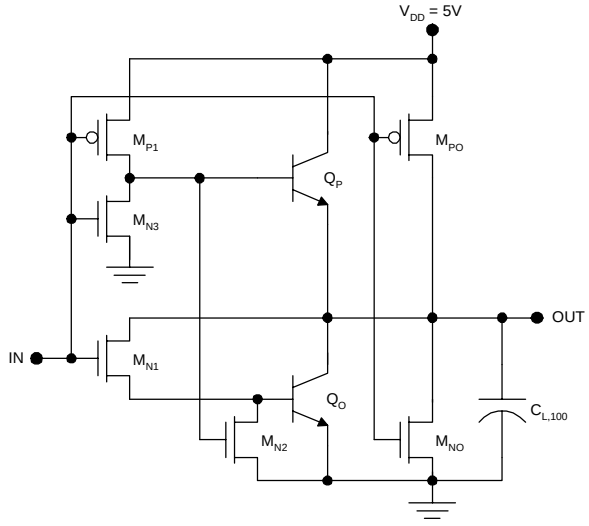


FIGURE 11.20
Rail-to-rail BiCMOS inverter (L11.4).

1. Using SPICE, determine the load capacitance, $C_{L,100}$, for which the average propagation delay is 100 ns.
2. Build the inverter and apply the load capacitance determined in (1).
3. Apply a 2-MHz square wave input, with a peak to peak amplitude of 5 V and a DC offset of 2.5 V.
4. Measure and plot the slew rate (the slew rate is dV_{OUT}/dt) vs. V_{OUT} for the low-to-high transition.
5. Repeat (4) for the high-to-low transition.
6. Discuss the variation of the slew rate and the practical implications.

Problems

P11.1. Consider the BiCMOS inverter shown in Figure 11.21.
1. Determine the mode of operation for each of the transistors.
2. Determine the supply current I_{DD}.
3. Determine the value of V_{OUT}.

P11.2. Consider the BiCMOS inverter shown in Figure 11.22.
1. Determine the mode of operation for each of the transistors.
2. Determine the supply current I_{DD}.
3. Determine the value of V_{OUT}.

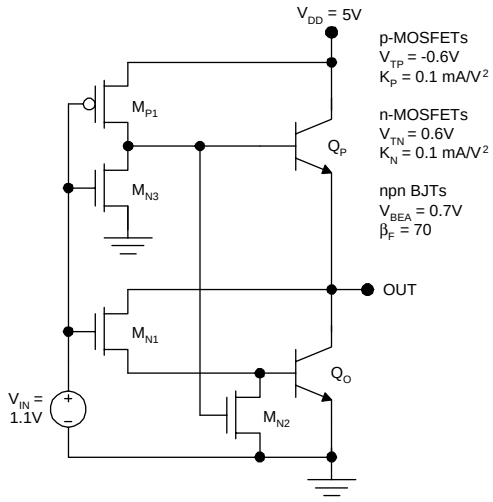


FIGURE 11.21
BiCMOS inverter (P11.1).

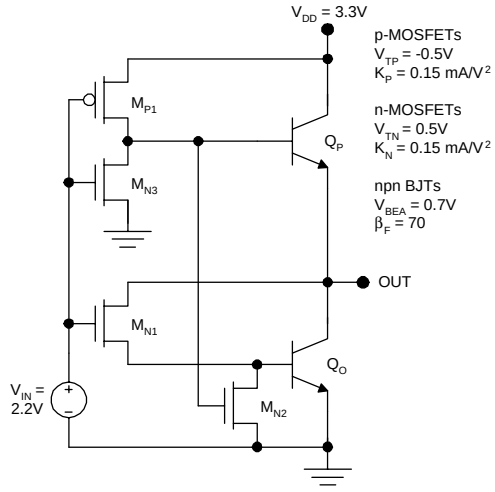


FIGURE 11.22
BiCMOS inverter (P11.2).

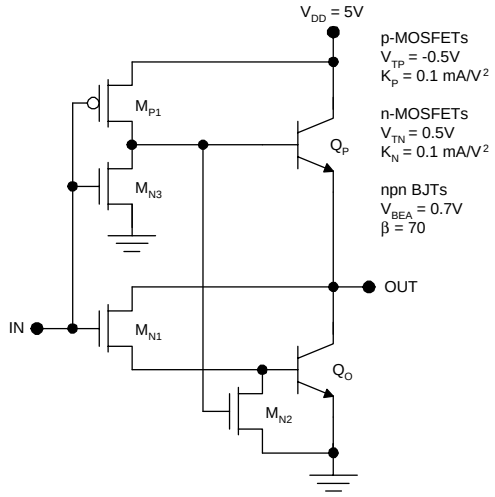


FIGURE 11.23
BiCMOS inverter (P11.3).

P11.3. Consider the BiCMOS inverter depicted in Figure 11.23.

1. Using hand calculations, determine and plot the voltage transfer characteristic.
2. Use SPICE to determine the voltage transfer characteristic.
3. Plot the two characteristics together; compare and contrast them.

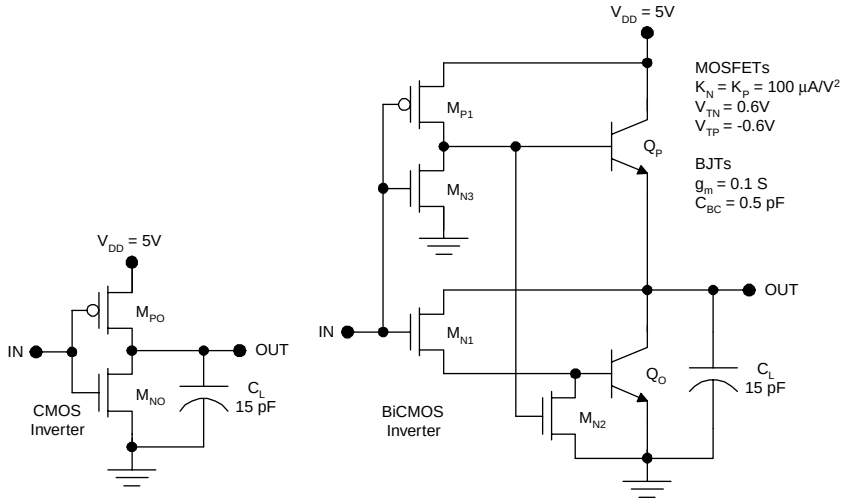


FIGURE 11.24
BiCMOS and CMOS gates (P11.4).

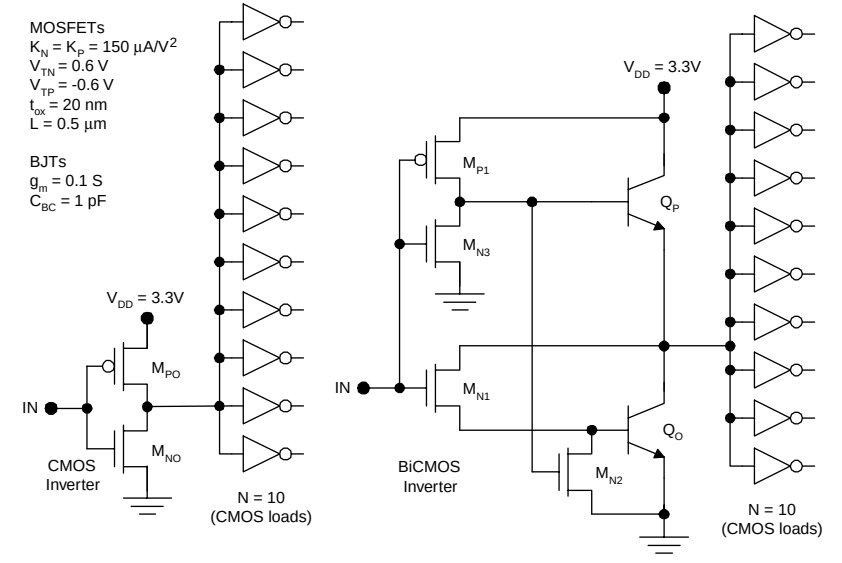


FIGURE 11.25
BiCMOS and CMOS gates (P11.5).

P11.4. For the BiCMOS and CMOS gates illustrated in Figure 11.24, compare the propagation delays for the case of 15-pF off-chip loads.

P11.5. For the BiCMOS and CMOS gates illustrated in Figure 11.25, compare the propagation delays for the case of 10 on-chip CMOS loads.

P11.6. Consider the BiCMOS and CMOS gates fabricated with the transistor parameters given in Table 11.4 through Table 11.6 and a supply voltage of V_{DD} = 5 V. For all n-MOSFETs, L_N = 0.25 μm and W_N = 1 μm. For all p-MOSFETs, L_p = 0.25 μm and W_p = 2.5 μm. Determine the value of the load capacitance for which the BiCMOS and CMOS gates exhibit equal propagation delays.

P11.7. Consider the BiCMOS and CMOS gates fabricated with the transistor parameters given in Table 11.4 through Table 11.6 and a supply voltage of V_{DD} = 5 V. For all n-MOSFETs, L_N = 0.25 μm and W_N = 1 μm. For all p-MOSFETs, L_p = 0.25 μm and W_p = 2.5 μm.

1. Determine and plot t_P vs. C_L for each type of circuit.
2. Determine the slopes of the characteristics in ohms.
3. Comment on the significance of the slopes.

P11.8. Consider a rail-to-rail BiCMOS gate with a CMOS shunted output, as shown in Figure 11.26 and the device parameters provided in Table 11.4 through Table 11.6. For all n-MOSFETs, L_N = 0.25 μm and W_N = 1 μm. For all p-MOSFETs, L_p = 0.25 μm and W_p = 2.5 μm.

TABLE 11.4

SPICE Parameters for the p-MOSFET for Problems

VTO	–0.6	CGDO	2.6p	TPG	1
KP	0.72E-4	CGBO	0	XJ	0
Gamma	1E-9	RSH	0	LD	0
Phi	6E-1	CJ	0	UO	230
Lambda	0.02	MJ	0.5	UCRIT	1E4
RD	0	CJSW	0	VMAX	1E5
RS	0	MJSW	0.33	NEFF	1
CBD	0	JS	0	KF	0
CBS	0	TOX	1E-8	AF	1
IS	1E-16	NSUB	1E15	FC	0.5
PB	8E-1	NSS	0	TNOM	27
CGSO	5.1p	NFS	0	Delta	0

TABLE 11.5

SPICE Parameters for the n-MOSFET for Problems

VTO	0.6	CGDO	2.6p	TPG	1
KP	1.84E-4	CGBO	0	XJ	0
Gamma	1E-9	RSH	0	LD	0
Phi	6E-1	CJ	0	UO	580
Lambda	0.02	MJ	0.5	UCRIT	1E4
RD	0	CJSW	0	VMAX	1E5
RS	0	MJSW	0.33	NEFF	1
CBD	0	JS	0	KF	0
CBS	0	TOX	1E-8	AF	1
IS	1E-16	NSUB	1E15	FC	0.5
PB	8E-1	NSS	0	TNOM	27
CGSO	5.1p	NFS	0	Delta	0

TABLE 11.6

SPICE npn BJT Parameters for Problems

IS	2E-15	ISC	0	TF	0.2n	CJS	0
BF	70	NC	2	XTF	0	VJS	0.75
NF	1	RB	0	VTF	1E100	MJS	0
VAF	1E100	IRB	1E100	ITF	0	XTB	0
IKF	1E100	RBM	0	PTF	0	EG	1.11
ISE	0	RE	0	CJC	0.15p	XTI	3
NE	1	RC	0	VJC	0.75	KF	0
BR	1	CJE	0.3p	MJC	0.5	AF	1
NR	1	VJE	0.8	XCJC	1	FC	0.5
VAR	1E100	MJE	0.333	TR	10n	TNOM	27
IKR	1E100						

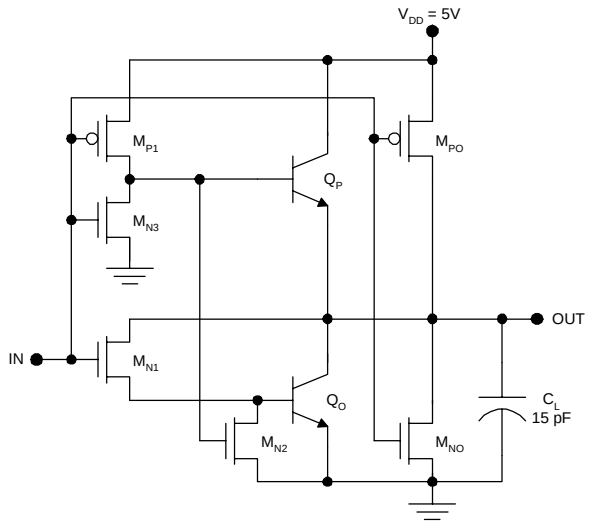


FIGURE 11.26
Rail-to-rail BiCMOS gate with a CMOS shunted output (P11.8).

1. Using SPICE, determine the propagation delays.
2. Using SPICE, apply a square wave input with a period equal to 10 times the average propagation delay, with a peak-to-peak amplitude of 5 V and a DC offset of 2.5 V. Determine and plot the slew rate (the slew rate is dV_{OUT}/dt) vs. V_{OUT} for the low-to-high transition.
3. Repeat (2) for the high-to-low transition.
4. Discuss the variation of the slew rate and the practical implications.

References

1. Harame, D., High performance BiCMOS process integration: trends, issues, and future directions, *Proc. 1997 Bipolar/BiCMOS Circuits Technol. Meet*., 36, 1997.
2. Kinoshita, Y., Suzuki, H., Nakamura, S., Fukaishi, M., Tajima, A., Sucmura, Y., Itani, T., Miyamoto, H., Fujii, H., Yotsuyanagi, M., Henmi, N., and Yamazaki, T., An advanced 0.25-µm BiCMOS process integration technology for multi-GHz communication LSIs, Proc. *1997 Bipolar/BiCMOS Circuits Technol. Meet*., 72, 1997.
3. Chiu, T.-Y. and Swartz, R.G., High performance BiCMOS for digital/analog applications, *Proc. 1992 IEEE Int. Symp. Circuits Syst*., 6, 3025, 1992.
4. Raje, P., Design and scaling of BiCMOS circuits, *Proc. 1992 IEEE Int. Conf. Computer Design: VLSI Computers Processors*, 234, 1992.
5. Shahidi, G.G., Warnock, J., Davari, B., Wu, B., Taur, Y., Wong, C., Chen, C.L., Rodriguez, M., Tang, D.D., Jenkins, K.A., McFarland, P.A., Schulz, R., Zicherman, D., Coane, P., Klaus, D., Sun, J.Y.C., Polcari, M., and Ning, T.H., A high performance BiCMOS technology using 0.25-µm CMOS and double poly 47 GHz bipolar, *Dig. Tech. Papers, 1992 Symp. VLSI Technol*., 28, 1992.
6. Watanabe, A., Nagano, T., Shukuri, S., and Ikeda, T., Future BiCMOS technology for scaled supply voltage, *Tech. Dig. 1989 Int. Electron. Devices Meet*., 429, 1989.
7. Wada, S., Nonaka, Y., Saito, T., Tominari, T., Koyu, K., Ikeda, K., Sakai, K., Sasahara, K., Watanabe, K., Fujiwara, H., Murata, F., Ohue, E., Kiyota, Y., Shimamoto, H., Washio, K., Takeyari, R., Hosoe, H., and Hashimoto, T., A manufacturable 0.18-µm SiGe BiCMOS technology for 40-Gb/s optical communication LSIs, *Proc. 2002 Bipolar/BiCMOS Circuits Technol. Meet*., 84, 2002.
8. Harame, D.L., Ahlgren, D.C., Coolbaugh, J.S. Dunn, J.S., Freeman, G.G., Gillis, J.D., Groves, R.A., Henderson, G.N., Johnson, A.J., Joseph, A.J., Subbanna, S., Victor, A.M., Watson, K.M., Webster, C.S., and Zampardi, P.J., Current status and future trends of SiGe BiCMOS technology, *IEEE Trans. Electron. Dev.*, 48, 2575, 2001.
9. www.intel.com (Intel).
10. www.motorola.com (Motorola).
11. www.amd.com (Advanced Micro Devices).
12. www.ibm.com (IBM).
13. www.lucent.com (Lucent Technologies)
14. www.fairchildsemi.com (Fairchild Semiconductor).
15. www.national.com (National Semiconductor).
16. www.cadence.com (Cadence).

12

GaAs Direct-Coupled FET Logic

12.1 Introduction

Gallium arsenide digital circuits hold the potential for higher performance than silicon devices because of the higher electron mobility in GaAs. However, no viable technology exists for the fabrication of GaAs MOSFETs. Because of the confluence of these circumstances, GaAs digital integrated circuits have been made using metal–semiconductor field-effect transistors (MESFETs). A number of circuit families have emerged but the most important is direct-coupled FET logic (DCFL).[1–18]

For a given lithographic technology, GaAs DCFL exhibits up to a threefold speed advantage over Si CMOS. However, this added speed comes at a cost, for two reasons: 1) GaAs wafers are smaller and more expensive, and 2) GaAs fabrication technology is more difficult and less mature than silicon technology. Also, as a consequence of the less mature GaAs processing, GaAs chips necessarily have fewer gates and less embedded memory than their silicon counterparts.

Based on the rapid progress in silicon CMOS, the threefold speed advantage realized with GaAs amounts to less than a 5-year lead. Due to higher cost and lower functionality, GaAs DCFL has succeeded only in the niches of high-end processors and high-speed communication.[1–8] In order to compete directly with CMOS, nonsilicon technologies will probably need to deliver a speed advantage of 10 times or more.

12.2 Gallium Arsenide vs. Silicon

The attractiveness of GaAs stems from the superior low-field mobility of electrons in this material. GaAs exhibits about six times the electron mobility of silicon for typical doping concentrations used in FETs as shown in Figure 12.1.[19] This translates into lower parasitic resistances and higher performance.

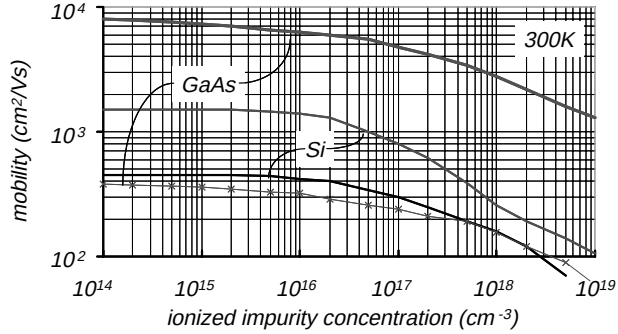


FIGURE 12.1
Electron and hole mobilities in GaAs and Si vs. the ionized impurity (doping) concentration.

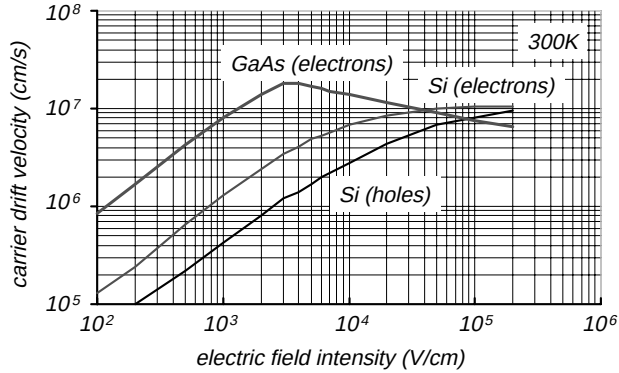


FIGURE 12.2
Carrier velocities vs. the electric field intensity in GaAs and Si.

Another benefit of the large low-field mobility is that the electrons in GaAs reach their maximum velocity at lower electric field strength; this is shown in Figure 12.2.[19] This means that a lower supply voltage can be used without a penalty in speed. This contrasts with silicon, in which the reduction of the supply voltage degrades the switching speed. As a result, GaAs circuits exhibit lower power delay products than their silicon competitors. Another fundamental advantage of GaAs is that the peak electron velocity is nearly twice that in silicon. For short channel FETs, the carriers move at close to the peak velocity. Therefore, a higher peak velocity translates into shorter times for the source to drain transit. This in turn increases the upper limit on the speed imposed by the transit time.

Despite the compelling case for GaAs, this material has three important drawbacks. First, there is no viable MOS technology; instead, MESFETs are used, but enhancement-type (normally off) devices are difficult to fabricate reproducibly. Second, holes in GaAs exhibit *lower* mobility than holes in silicon, precluding the development of complementary n-channel and p-channel transistors. Instead, circuits similar to NMOS are used at the

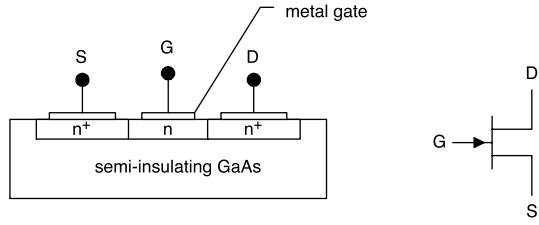


FIGURE 12.3
GaAs MESFET: schematic and circuit symbol.

expense of high standby dissipation. Third, GaAs circuits are much more costly to fabricate than silicon ICs because silicon wafers are four times larger in diameter, one tenth the cost per square centimeter, easier to process, and typically produce a higher yield of working die.

12.3 Gallium Arsenide MESFET

The metal–semiconductor field-effect transistor, or MESFET, is a majority carrier (drift) three-terminal device. The three terminals are the source, drain, and gate. The device takes its name from the fact that the gate forms a metal–semiconductor (Schottky) diode. The device structure is shown schematically in Figure 12.3.

With zero gate–source bias, a depletion region is in the n-type channel region under the metal gate. However, unless this depletion layer extends all the way through the channel layer, a conducting n-type channel will still exist between the source and drain. Therefore the MESFET is usually a "normally on" or depletion type transistor. Variation of the gate–source bias voltage will modulate the depletion width under the gate and therefore the conductivity of the channel. If a sufficiently large reverse bias is applied to the gate junction, the gate depletion region will extend through the channel and the channel will "pinch off." The gate-source voltage necessary to cause pinch-off is called the *pinch-off voltage*.

12.4 Metal–Semiconductor Junction

Invariably MESFETs are fabricated using an n-type semiconductor. The zero-bias energy band diagram for the resulting metal–semiconductor junction is shown in Figure 12.4. With zero bias applied, the Fermi levels in the metal and semiconductor line up. The difference between the Fermi level in the metal and the vacuum level is the metal work function $q\phi_m$. (This is the

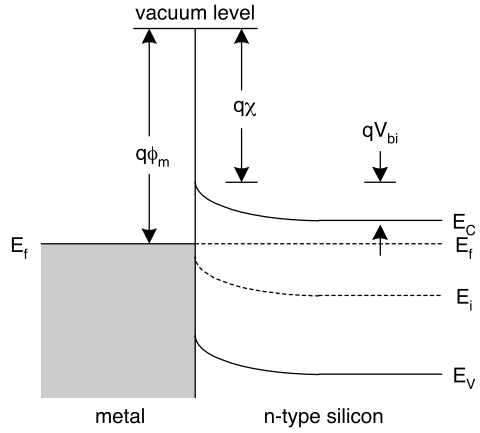


FIGURE 12.4
Zero-bias energy band diagram for a metal–n-semiconductor junction.

energy necessary to remove an electron from the metal to a vacuum.) The difference between the semiconductor conduction band and the vacuum level is the semiconductor electron affinity $q\chi$. The built-in voltage for the metal–semiconductor junction is given by

$$V_{bi} = q\phi_m - q\chi - \frac{kT}{q}\ln\left(\frac{N_C}{N_d}\right), \tag{12.1}$$

where N_C is the effective density of states in the conduction band and N_d is the donor concentration in the semiconductor.

No depletion region exists in the metal. In the n-type semiconductor, the zero-bias depletion width can be determined by solution of Poison's equation in much the same way as for a p–n junction. It is

$$h = \sqrt{\frac{2\varepsilon_s V_{bi}}{qN_d}}\,. \tag{12.2}$$

With a bias V applied between the metal and the semiconductor, the depletion width is modified to

$$h = \sqrt{\frac{2\varepsilon_s\left(V_{bi} - V\right)}{qN_d}}\,. \tag{12.3}$$

12.5 MESFET Pinch-Off Voltage

Consider a MESFET with the dimensions indicated in Figure 12.5. With zero drain–source bias, the depleted height in the channel is the same at the source

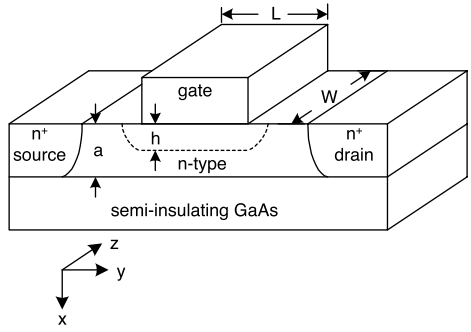


FIGURE 12.5
GaAs n-channel MESFET with zero drain–source bias.

and drain ends, as shown in this figure. The channel of the MESFET pinches off when the depletion width is equal to the channel height a:

$$a = h\big|_{V_{GS}=V_P}; \tag{12.4}$$

therefore, the pinch-off voltage is given by

$$V_P = V_{bi} - \frac{qN_d a^2}{2\varepsilon_s}.$$

For a depletion-type MESFET the pinch-off voltage is negative, but for an enhancement-type device the pinch-off voltage is positive.

Example 12.1

Calculate the pinch-off voltage for a GaAs MESFET with V_{bi} = 1.0 V, N_d = 2×10^{17} cm^{-3}, and a = 0.15 μm.

Solution. The pinch-off voltage is

$$V_P = V_{bi} - \frac{qN_d a^2}{2\varepsilon}$$

$$= 1.0\text{ V} - \frac{\left(1.6\times10^{-19}\text{ C}\right)\left(2\times10^{17}\text{ cm}^{-3}\right)\left(0.15\times10^{-4}\text{ cm}\right)^2}{2(13.1)\left(8.85\times10^{-14}\text{ F/cm}\right)} = -2.1\text{ V}$$

The negative pinch-off voltage indicates that this is a depletion type (normally on) device.

12.6 Long-Channel MESFET Operation

Consider an n-channel depletion-type MESFET with a gate–source bias V_{GS} and a drain–source bias V_{DS}, with three modes of operation for the MESFET: *cutoff*, *linear*, and *saturation*. *Cutoff* occurs if the gate-to-source bias voltage is more negative than the pinch-off voltage. As a first-order approximation, cutoff results in zero drain current. If the gate-to-source bias is made more positive than the pinch-off voltage for the device, then a conducting channel is induced and a drain current can flow. With a small drain-to-source bias, the MESFET acts like a voltage-controlled resistance; this is the *linear* (also known as ohmic or triode) mode of operation. However, if the drain-to-source bias is sufficiently large, then the conducting channel will pinch off at the drain end. This causes the drain current to saturate, so this mode of operation is called *saturation*.

12.6.1 MESFET Cutoff Operation

Cutoff operation occurs if the gate-to-source bias is more negative than the pinch-off voltage. To a first approximation, cutoff is accompanied with zero drain current:

$$I_D \approx 0 \quad \left(V_{GS} < V_P\right). \tag{12.5}$$

12.6.2 MESFET Linear Operation

Linear operation occurs if the gate-to-source bias is more positive than the pinch off and the drain-to-source bias is small enough so that the channel does not pinch off at the drain end. In the linear mode of operation, the MESFET acts like a voltage-controlled resistance. The undepleted channel will be x_1 at the source end and x_2 at the drain end as shown in Figure 12.6.

At a point y along the channel, the depletion width is

$$h(y) = \left[\frac{2\varepsilon_s\left(V(y) + V_{bi} - V_{GS}\right)}{qN_d}\right]^{1/2}, \tag{12.6}$$

where $V(y)$ is the potential in the channel. At the source end, $V(y) = 0$, so

$$x_1 = \left[\frac{2\varepsilon_s\left(V_{bi} - V_{GS}\right)}{qN_d}\right]^{1/2}. \tag{12.7}$$

At the drain end, $V(y) = V_{DS}$, so

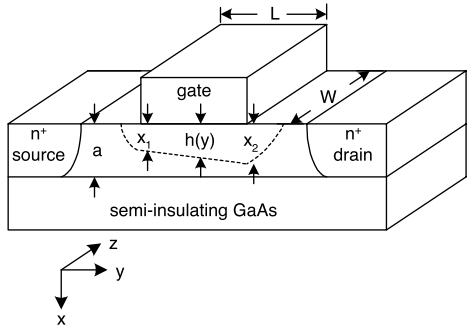


FIGURE 12.6
MESFET with a gate–source bias V_{GS} and drain–source bias V_{DS}.

$$x_2 = \left[\frac{2\varepsilon_s\left(V_{DS} - V_{bi} - V_{GS}\right)}{qN_d}\right]^{1/2}. \tag{12.8}$$

Assuming the current flow is one dimensional (the *gradual channel approximation*), then

$$I_D = q\mu_n N_d\left[a - h(y)\right]W\frac{dV}{dx}; \tag{12.9}$$

however, from Equation 12.6,

$$dV\frac{qhN_d}{\varepsilon_s}dh, \tag{12.10}$$

so

$$I_D dy = \frac{q^2 N_d^2 \mu_n}{\varepsilon_s}\left[a - h(y)\right]h(y)Wdh. \tag{12.11}$$

The drain current in the linear mode of operation can be determined by integrating over the channel length.

$$\begin{aligned} I_D &= \frac{1}{L}\int_{y_1}^{y_2}\frac{q^2 N_d^2 \mu_n}{\varepsilon_s}\left[a - h(y)\right]Wdh \\ &= \frac{Wq^2 N_d^2 \mu_n}{L\varepsilon_s}\left[\frac{ah^2}{2} - \frac{h^3}{3}\right]_{y_1}^{y_2} \\ &= I_P\left[\frac{3V_{DS}}{V_{bi} - V_P} - \frac{2\left(V_{DS} + V_{bi} - V_{GS}\right)^{3/2} - 2\left(V_{bi} - V_{GS}\right)^{3/2}}{\left(V_{bi} - V_P\right)^{3/2}}\right], \end{aligned} \tag{12.12}$$

where

$$I_P = \frac{Wq^2 N_d^2 a^3 \mu_n}{6L\varepsilon_s}. \tag{12.13}$$

12.6.3 MESFET Saturation Operation

Saturation operation occurs if the drain-to-source bias is sufficient to cause the channel to pinch off at the drain end. This occurs when

$$V_{DS} \geq V_{GS} - V_P. \tag{12.14}$$

Substituting $V_{DS} = V_{GS} - V_P$ in the equation for linear operation yields the equation for saturated drain current in the MESFET:

$$I_D = I_P \left[\frac{3\left(V_{GS} - V_P\right)}{V_{bi} - V_P} - \frac{2\left(V_{bi} - V_P\right)^{3/2} - 2\left(V_{bi} - V_{GS}\right)^{3/2}}{\left(V_{bi} - V_P\right)^{3/2}} \right]. \tag{12.15}$$

12.6.4 Transit Time

It takes a finite time for majority carriers to traverse the channel in a conducting MESFET, called the transit time. This is very similar to the case of the MOSFET. As with MOSFETs, the assumption is usually that the transit time is much shorter than the circuit delays; this is called the quasi-static assumption. In a long-channel MESFET, the drift of electrons in the channel is governed by Ohm's law and the low-field mobility, so the transit time is given by

$$t_t = \frac{L^2}{\mu_n V_{DS}}. \tag{12.16}$$

Therefore, the transit time increases with the square of the channel length.

Example 12.2

Estimate the transit time of a GaAs MESFET with a 2-μm channel length in a 1-V DCFL circuit.

Solution. Assuming a room-temperature electron mobility of 3000 cm^2/Vs, the transit time is

$$t_t = \frac{L^2}{\mu_n V_{DS}} = \frac{\left(2 \times 10^{-4}\ \text{cm}\right)^2}{\left(3000\ \text{cm}^2\text{V}^{-1}\text{s}^{-1}\right)(1\ \text{V})} = 13\ \text{ps}.$$

The quasi-static approximation is applicable if the circuit propagation delay is greater than the transit time.

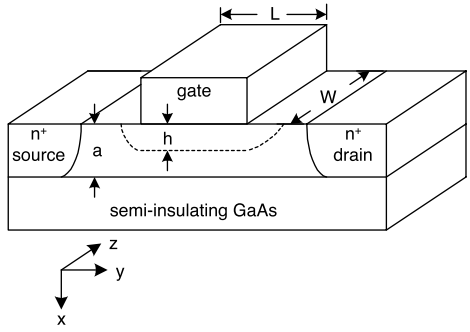


FIGURE 12.7
GaAs MESFET for consideration of the current in the limiting case of electron velocity saturation.

12.7 Short-Channel MESFETs

Rapid advances in lithography and the resulting scaling of channel lengths in MESFETs have resulted in devices with different electrical behavior compared to long-channel devices. First, the electric field intensity in the channel may be sufficiently large so that it is no longer valid to use Ohm's law and the low-field electron mobility. It is possible for electrons in the channel to move at their saturation velocity for part or most of the length of the channel. Second, the effective channel length becomes a function of the drain-to-source bias as a consequence of *channel length modulation*. These effects must be considered in the design of high-performance DCFL circuits today.

12.7.1 Field-Dependent Mobility

At high electric field intensities, the carrier drift velocities are no longer proportional to the electric field. Instead, there is velocity overshoot and then saturation as illustrated in Figure 12.2. Consider the limiting case in which electrons move at their saturated velocity over the entire length of the channel; the undepleted channel height is constant over the channel length as shown in Figure 12.7. In this case, the drain current for saturation operation of the device under the assumption of carrier velocity saturation is

$$I_D = qN_d v_{sat} W(a-h). \tag{12.17}$$

Therefore, the drain current for saturation operation is less than that for the constant mobility case. Also, the saturated drain current becomes a function of the device width alone and not the aspect ratio W/L.

In practical devices, the carriers may move at the saturated velocity over a portion, but not all, of the channel length. This may be modeled by a two-region model comprising a constant mobility MESFET in series with a saturated velocity MESFET. Fortunately, the resulting device characteristics will be bracketed between the two limiting cases of constant mobility and velocity saturation. An important exception to this rule occurs in the case of ballistic transport, for which the electrons traverse the channel without colliding with any source of scattering. The scattering centers are responsible for determining the bulk velocity vs. field characteristic, which does not apply to ballistic transport.

12.7.2 Transit Time in Short-Channel MESFETs

In short-channel MESFETs, the carriers may travel at close to the saturation velocity for the entire length of the channel. In this limit the transit time is

$$t_t = \frac{L}{v_{sat}} . \tag{12.18}$$

Therefore the transit time is directly proportional to the channel length (rather than the square of the channel length) in short-channel MESFETs.

12.7.3 Channel Length Modulation

The drain current in a MESFET saturates at the value of V_{DS}, which causes the channel to pinch off at the drain end. Further increase in V_{DS} causes the pinch-off point to move into the channel toward the source, as shown in Figure 12.8. This increases the drain current by the ratio $L/(L - \Delta L)$. In a long-channel MESFET, this effect is of little consequence because the percentage change in the drain current is small. However, the *channel length modulation effect* is important in short-channel MESFETs. Mathematically, the channel length modulation is modeled by multiplying the drain current expressions by a factor that increases linearly with the drain-to-source bias, in the same manner as for MOSFETs.

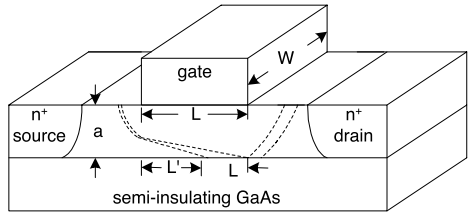


FIGURE 12.8
Channel length modulation in a short-channel MOSFET.

12.8 The Curtice Model for the MESFET

Because of the complexity of the analytical models for the MESFET, it is common to use the empirical Curtice model.[20,21] This model uses a single equation for linear and saturated operations of the MESFET and is given by

$$I_D = \beta\left(V_{GS} - V_P\right)^2\left(1 + \lambda V_{DS}\right)\tanh\left(\alpha V_{DS}\right); \qquad \left[V_{GS} \geq V_P\right], \tag{12.19}$$

where β = the device transconductance parameter, given by

$$\beta = \frac{\mu_n \varepsilon_s}{2a}\frac{W}{L}, \tag{12.20}$$

λ = the channel length modulation parameter, and α = the tanh parameter.

The preceding model holds for the constant mobility case and is applicable to long-channel MESFETs. In short-channel MESFETs, field-dependent mobility must be considered. One approach here is to use a modified value of the device transconductance parameter,

$$\beta' = \frac{2\mu_n \varepsilon_s v_{sat} W}{a\left[\mu_n\left(\left|V_P\right| + V_{bi}\right) + 4v_{sat}L\right]}. \tag{12.21}$$

In practice, the field-dependent mobility must be considered for MESFETs with channel lengths less than 1.0 μm.

The Curtice model is used for long- and short-channel MESFETs in SPICE because of its reasonable accuracy and computational simplicity. However, the parameters β, α, and λ are usually determined empirically.

Example 12.3

Consider a long-channel GaAs MESFET with V_{bi} = 1.0 V, $N_d = 2 \times 10^{17}$ cm^{-3}, a = 0.15 μm, L = 1.0 μm, and W = 2.0 μm. Calculate and plot the characteristic curves using the analytical model. Repeat using the Curtice model assuming α = 1.5 V^{-1} and λ = 0.01 V^{-1}.

Solution. As calculated in Example 12.1, V_p = –2.1 V. Assuming long-channel behavior and constant mobility (μ_n = 3000 cm^{2}V^{-1}s^{-1}),

$$I_P = \frac{Wq^2N_d^2a^3\mu_n}{6L\varepsilon_s}$$

$$= \frac{\left(2\times10^{-4}\text{ cm}\right)\left(1.6\times10^{-19}\right)^2\left(2\times10^{17}\text{ cm}^{-3}\right)^2\left(0.15\times10^{-4}\text{ cm}\right)^3\left(3000\text{ cm}^2\text{V}^{-1}\text{s}^{-1}\right)}{6\left(10^{-4}\text{ cm}\right)(13.1)\left(8.85\times10^{-14}\text{ F/cm}\right)}.$$

$$= 3.0\text{ mA}$$

Using the analytical model, the device characteristics are given by

$$I_D = \begin{cases} 0, & V_{GS} \le V_P; \\ I_P\left[\dfrac{3V_{DS}}{V_{bi}-V_P} - \dfrac{2\left(V_{DS}+V_{bi}-V_{GS}\right)^{3/2} - 2\left(V_{bi}-V_{GS}\right)^{3/2}}{\left(V_{bi}-V_P\right)^{3/2}}\right], & V_{DS} \le V_{GS}-V_P; \text{ and,} \\ I_P\left[\dfrac{3\left(V_{GS}-V_P\right)}{V_{bi}-V_P} - \dfrac{2\left(V_{bi}-V_P\right)^{3/2} - 2\left(V_{bi}-V_{GS}\right)^{3/2}}{\left(V_{bi}-V_P\right)^{3/2}}\right], & V_{DS} \ge V_{GS}-V_P. \end{cases}$$

For the example MESFET,

$$I_D = \begin{cases} 0, & V_{GS} \le -2.1\text{ V}; \\ 3.0\text{ mA}\left[\dfrac{3V_{DS}}{3.1V} - \dfrac{2\left(V_{DS}+1.0\text{ V}-V_{GS}\right)^{3/2} - 2\left(1.0\text{ V}-V_{GS}\right)^{3/2}}{\left(3.1\text{ V}\right)^{3/2}}\right], & \\ \qquad V_{DS} \le V_{GS}+2.1\text{ V}; \text{ and} & \\ 3.0\text{ mA}\left[\dfrac{3\left(V_{GS}+2.1\text{ V}\right)}{3.1\text{ V}} - \dfrac{2\left(3.1\text{ V}\right)^{3/2} - 2\left(1.0\text{ V}-V_{GS}\right)^{3/2}}{\left(3.1\text{ V}\right)^{3/2}}\right], & \\ \qquad V_{DS} \ge V_{GS}+2.1\text{ V}. & \end{cases}$$

For application of the Curtice model, the device transconductance parameter is

$$\beta = \frac{\mu_n \varepsilon_s}{2a}\frac{W}{L} = \frac{\left(3000\text{ cm}^2\text{V}^{-1}\text{s}^{-1}\right)(13.1)\left(8.85\times10^{-14}\text{ F/cm}\right)}{2\left(0.15\times10^{-4}\text{ cm}\right)}\frac{2\times10^{-4}\text{ cm}}{10^{-4}\text{ cm}}$$

$$= 0.23\text{ mA/V}^2.$$

Using the Curtice model, the characteristics are given by

$$I_D = \begin{cases} 0; & V_{GS} \le -2.1\text{ V} \\ 0.23\text{ mA/V}^2\left(V_{GS}+2.1\text{ V}\right)^2\left(1+0.01V_{DS}\right)\tanh\left(1.5V_{DS}\right); & V_{GS} \ge -2.1\text{ V} \end{cases}$$

Both sets of characteristics are plotted in Figure 12.9. The maximum departure between the analytical characteristics and the Curtice model is less than 20%.

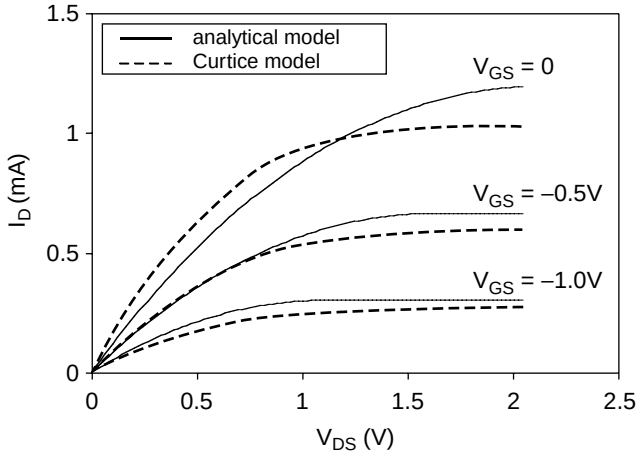


FIGURE 12.9
Example characteristics for a GaAs MESFET with V_{bi} = 1.0 V, $N_d = 2 \times 10^{17}$ cm^{-3}, a = 0.15 μm, L = 1.0 μm, and W = 2.0 μm. The pinch-off voltage is –2.1 V. For the purpose of calculating characteristics with the Curtice model, it was assumed that $\alpha = 1.5$ V^{-1} and $\lambda = 0.01$ V^{-1}.

Example 12.4

Consider a GaAs MESFET with V_{bi} = 1.0 V, $N_d = 2 \times 10^{17}$ cm^{-3}, a = 0.15 μm, L = 0.25 μm, and W = 1.0 μm. Estimate the device transconductance parameter assuming long-channel behavior with constant mobility ($\mu_n = 3000$ cm^{2}V^{-1}s^{-1}). Repeat, assuming short-channel behavior (use the value of β′ to account for the field-dependent mobility and assume a saturation velocity of 10^7 cm/s).

Solution. Assuming long-channel behavior and constant mobility (μ_n = 3000 cm^{2}V^{-1}s^{-1}), the device transconductance parameter is

$$\beta = \frac{\mu_n \varepsilon_s}{2a}\frac{W}{L} = \frac{\left(3000\ \text{cm}^2\text{V}^{-1}\text{s}^{-1}\right)(13.1)\left(8.85\times10^{-14}\ \text{F/cm}\right)}{2\left(0.15\times10^{-4}\ \text{cm}\right)}\frac{10^{-4}\ \text{cm}}{0.25\times10^{-4}\ \text{cm}}.$$

$$= 0.46\ \text{mA/V}^2$$

Assuming short-channel behavior, a low-field mobility μ_n = 3000 cm^{2}V^{-1}s^{-1}, and an electron saturation velocity of 10^7 cm/s, the device transconductance parameter is

$$\beta' = \frac{2\mu_n \varepsilon_s v_{sat} W}{a\left[\mu_n\left(|V_P| + V_{bi}\right) + 4v_{sat}L\right]}$$

$$= \frac{2\left(3000\ \text{cm}^2\text{V}^{-1}\text{s}^{-1}\right)(13.1)\left(8.85\times10^{-14}\ \text{F/cm}\right)\left(10^7\ \text{cm/s}\right)\left(10^{-4}\ \text{cm}\right)}{\left(0.15\times10^{-4}\ \text{cm}\right)\left[\left(3000\ \text{cm}^2\text{V}^{-1}\text{s}^{-1}\right)(2.1V + 1\ \text{V}) + 4\left(10^7\ \text{cm/s}\right)\left(0.5\times10^{-4}\ \text{cm}\right)\right]}.$$

$$= 0.041\ \text{mA}$$

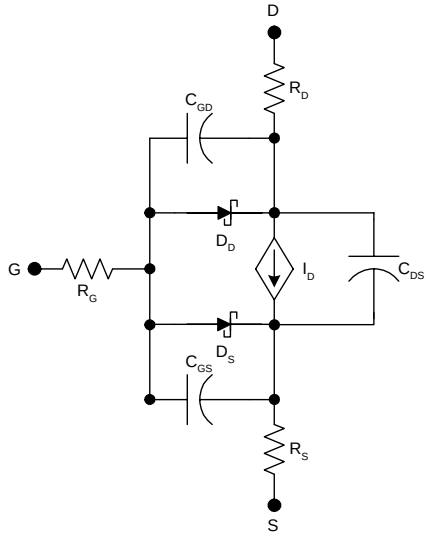


FIGURE 12.10
SPICE MESFET model.

Here, the effective device transconductance parameter has been reduced by an order of magnitude as a consequence of the field-dependent mobility. Clearly, the constant mobility model is not appropriate for a 0.25-μm GaAs MESFET. In practice, field-dependent mobility must be considered for submicron GaAs MESFETs.

12.9 MESFET SPICE Model

SPICE uses the MESFET model shown in Figure 12.10. I_D is a Curtice type current source and R_D, R_S, and R_G are the parasitic resistances in the drain, source and gate, respectively. C_{GD} is the capacitance between the gate and drain, C_{GS} is the capacitance between the gate and the source, and C_{DS} is the drain source capacitance. The diodes are parasitic metal–semiconductor junctions between the body and the channel; their forward bias must be limited in order to minimize the DC gate current.

SPICE models the MESFET current source using the Curtice model:

$$I_D = BETA\left(V_{GS} - VTO\right)^2\left(1 + LAMBDAV_{DS}\right)\tanh\left(ALPHAV_{DS}\right). \quad (12.22)$$

The diodes are modeled using Shockley type sources,

TABLE 12.1

SPICE Parameters for the GaAs MESFET

Symbol	SPICE Name	Description	Units	Default	Typical
V_P	VTO	Pinch-off voltage	V	–2.5	—
β	BETA	Device transconductance parameter	A/V	0.1	0.1
λ	LAMBDA	Channel length modulation parameter	V^{-1}	0	1E-4
α	ALPHA	Saturation parameter	V^{-1}	2.0	2.0
I_S	IS	Gate saturation current	A	1E-14	1E-14
N	N	Gate emission current	—	1	1.5
R_G	RG	Gate resistance	Ω	0	5
R_D	RD	Drain resistance	Ω	0	1
R_S	RS	Source resistance	Ω	0	1
C_{GS}	CGS	Gate-to-source capacitance with zero bias	F	0	1E-13
C_{GD}	CGD	Gate-to-drain capacitance with zero bias	F	0	1E-13
C_{DS}	CDS	Drain-to-source capacitance with zero bias	F	0	1E-14

$$I_{DD} = IS\left[\exp\left(\frac{V_{GD}}{NV_T}\right) - 1\right], \text{ and} \tag{12.23}$$

$$I_{DS} = IS\left[\exp\left(\frac{V_{GS}}{NV_T}\right) - 1\right]. \tag{12.24}$$

where

V_{GS} = gate-to-source voltage
V_{DS} = drain-to-source voltage
I_D = drain current source
V_{TO} = pinch-off voltage
BETA = device transconductance parameter
ALPHA = tanh parameter ("saturation parameter")
LAMBDA = channel length modulation parameter
IS = Schottky diode reverse saturation current
N = Schottky diode emission coefficient

The SPICE MESFET model parameters are listed in Table 12.1.

12.10 Integrated MESFETs

Integrated MESFETs are fabricated on semi-insulating GaAs substrates as shown in Figure 12.11. The source and drain regions are n^+-type and the

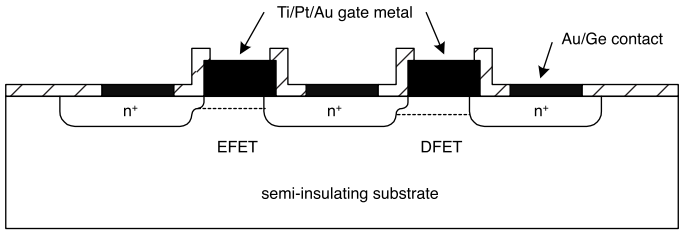


FIGURE 12.11
Integrated GaAs MESFETs on a semi-insulating substrate.

channel regions are n-type. Ion implantation gives the necessary control in doping to allow the fabrication of enhancement type and depletion type transistors on the same wafer (labeled EFET and DFET, respectively). The gate metal is typically Ti/Pt/Au and an Au/Ge eutectic alloy is used for the ohmic contacts. The devices are self-isolating as a result of the semi-insulating substrate, which typically exhibits a resistivity of 10^8 Ωcm.

12.11 Direct-Coupled FET Logic (DCFL)

DCFL is the most important gallium arsenide digital logic family. Early GaAs logic circuits were fabricated using only depletion type (normally on) MESFETs; DCFL uses depletion type and enhancement type devices. This improvement obviates the need for level-shifting circuitry or dual power supplies. A DCFL inverter is shown in Figure 12.12. The switch transistor, M_{NO}, is an EFET ($V_{PO} > 0$) and the load device M_{NL} is a DFET ($V_{PL} < 0$).

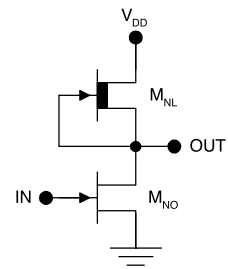


FIGURE 12.12
GaAs DCFL inverter.

The basic operation of the DCFL inverter is as follows. If the input voltage is less than V_{PO}, then M_{NO} is cut off and M_{NL} is linear; therefore, the output goes high (to V_{DD}). On the other hand, with a logic-one input, M_{NO} is linear while M_{NL} is saturated, causing the output voltage to go low (to V_{OL}). Thus the operation is very similar to that of the NMOS inverter, but with one critical difference: the input voltage cannot be so positive that it turns on the Schottky gate junction in M_{NO}. Thus, V_{OH} (and therefore V_{DD}) must be less than the turn-on voltage for the metal–semiconductor junction. Practically, this means that GaAs DCFL cannot operate with a supply voltage greater than about 1 V.

Because of the similarity between DCFL and NMOS circuitry, many of the electrical characteristics of GaAs DCFL may be predicted using the NMOS

theory. In applying NMOS equations to DCFL, K_O and V_{TO} are replaced with β_O and V_{PO}, respectively, and K_L and V_{TL} are replaced with β_L and V_{PL}, respectively. The resulting predictions are quite approximate because the characteristics of the MESFET and the MOSFET have different mathematical forms. However, the approximations are simple and provide factor-of-two accuracy. In addition, they clearly show how circuit and device design parameters affect the circuit performance.

12.11.1 Voltage Transfer Characteristic

If the Curtice model is used, only two regions of the voltage transfer characteristic need be considered. In the first region, the pull-down transistor is cut off and the pull-up transistor is linear so that

$$V_{\text{OUT}} = V_{DD}; \quad V_{\text{IN}} \le V_{PO}, \tag{12.25}$$

where V_{PO} is the pinch-off voltage for the pull-down transistor, M_{PO}. In the second region, both transistors conduct and the voltage transfer characteristic may be determined by equating the two drain currents and solving. Therefore, for $[V_{GS} \ge V_{PO}]$,

$$\begin{aligned}&\beta_O\left(V_{\text{IN}} - V_{PO}\right)^2\left(1+\lambda_O V_{\text{OUT}}\right)\tanh\left(\alpha V_{\text{OUT}}\right)\\ &\quad = \beta_L\left(-V_{PL}\right)^2\left[1+\lambda_L\left(V_{DD}-V_{\text{OUT}}\right)\right]\tanh\left[\alpha_L\left(V_{DD}-V_{\text{OUT}}\right)\right]\end{aligned} \tag{12.26}$$

where

β_O = device transconductance parameter for the pull-down device
V_{PO} = pinch-off voltage for the pull-down device
λ_O = channel length modulation parameter for the pull-down device
α_O = tanh parameter for the pull-down device
β_L = device transconductance for the pull-up device
V_{PL} = pinch-off voltage for the pull-up device
λ_L = channel length modulation parameter for the pull-up device
α_L = tanh parameter for the pull-up device

There is no analytical solution for V_{OUT}. However, the entire voltage transfer characteristic may be determined by calculating the values of V_{IN} corresponding to assumed values of V_{OUT}, using the relationship

$$V_{\text{IN}} = V_{PO} + \sqrt{\frac{\beta_L\left(-V_{PL}\right)^2\left[1+\lambda_L\left(V_{DD}-V_{\text{OUT}}\right)\right]\tanh\left[\alpha_L\left(V_{DD}-V_{\text{OUT}}\right)\right]}{\beta_O\left(1+\lambda_O V_{\text{OUT}}\right)\tanh\left(\alpha_O V_{\text{OUT}}\right)}}. \tag{12.27}$$

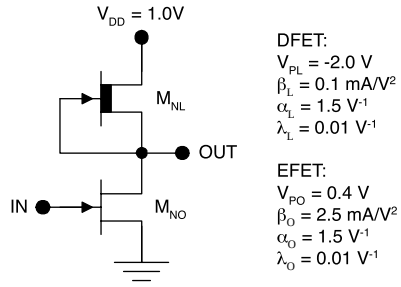


FIGURE 12.13
Example DCFL inverter for the calculation of the voltage transfer characteristic.

Using the NMOS analogy provides approximate equations for the critical voltages in the VTC, as follows:

$$V_{IL} \approx V_{PO} + \frac{\beta_L}{\sqrt{\beta_O \beta_L + \beta_O^2}} |V_{PL}|, \tag{12.28}$$

$$V_{IH} \approx V_{PO} + 2|V_{PL}| \sqrt{\frac{\beta_L}{3\beta_O}}, \tag{12.29}$$

and

$$V_{OL} \approx V_{DD} - V_{PO} - \sqrt{(V_{DD} - V_{PO})^2 - \left(\frac{\beta_L}{\beta_O}\right) V_{PL}^2}. \tag{12.30}$$

It should be noted that the preceding approximations yield a factor-of-two accuracy and do not account for the channel length modulation or saturation parameters of the MESFETs.

Example 12.5

Calculate the voltage transfer characteristic for the DCFL inverter shown in Figure 12.13 using Equation 12.27 and determine the critical voltages. Also estimate the critical voltages using approximate Equation 12.28 through Equation 12.30 for comparison.

Solution. The voltage transfer characteristic can be determined numerically using the relationship

$$V_{\text{IN}} = V_{PO} + \sqrt{\frac{\beta_L (-V_{PL})^2 \left[1 + \lambda_L (V_{DD} - V_{\text{OUT}})\right] \tanh\left[\alpha_L (V_{DD} - V_{\text{OUT}})\right]}{\beta_O (1 + \lambda_O V_{\text{OUT}}) \tanh(\alpha_O V_{\text{OUT}})}}$$

$$= 0.4 \text{ V} + \sqrt{\frac{0.1 \text{ mA/V}^2 (2 \text{ V})^2 \left[1 + 0.01 \text{ V}^{-1} (1 \text{ V} - V_{\text{OUT}})\right] \tanh\left[1.5 \text{ V}^{-1} (1 \text{ V} - V_{\text{OUT}})\right]}{2.5 \text{ mA/V}^2 \left(1 + 0.01 \text{ V}^{-1} V_{\text{OUT}}\right) \tanh\left(1.5 \text{ V}^{-1} V_{\text{OUT}}\right)}}.$$

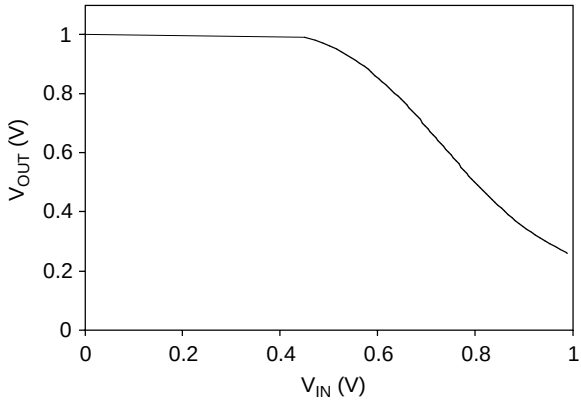


FIGURE 12.14
Voltage transfer characteristic for a GaAs DCFL inverter with V_{DD} = 1.0 V, β_L = 0.1 mA/V^2, α_L = 1.5 V^{-1}, λ_L = 0.01 V^{-1}, β_O = 2.5 mA/V^2, α_O = 1.5 V^{-1}, and λ_O = 0.01 V^{-1}.

The voltage transfer characteristic shown in Figure 12.14 was determined point by point, by assuming values of V_{OUT} and calculating the corresponding values of V_{IN}. The critical voltages can be determined from the numerical results as follows: V_{IL} = 0.45 V; V_{IH} = 0.89 V; V_{OH} = 1.0 V; and V_{OL} = 0.26 V.

Using the NMOS analogy provides approximate equations for the critical voltages in the VTC, as follows:

$$V_{IL} \approx V_{PO} + \frac{\beta_L}{\sqrt{\beta_O\beta_L + \beta_O^2}}|V_{PL}|$$

$$= 0.4\text{ V} + \frac{0.1\text{ mA/V}^2}{\sqrt{(2.5\text{ mA/V}^2)(0.1\text{ mA/V}^2) + (2.5\text{ mA/V}^2)^2}}|-2\text{ V}|,$$

$$= 0.48\text{ V}$$

$$V_{IH} \approx V_{PO} + 2|V_{PL}|\sqrt{\frac{\beta_L}{3\beta_O}}$$

$$= 0.4\text{ V} + 2|-2\text{ V}|\sqrt{\frac{0.1\text{ mA/V}^2}{3(2.5\text{ mA/V}^2)}},$$

$$= 0.86\text{ V}$$

and

$$V_{OL} \approx V_{DD} - V_{PO} - \sqrt{(V_{DD} - V_{PO})^2 - \left(\frac{\beta_L}{\beta_O}\right)V_{PL}^2}$$

$$= 1.0\text{ V} - 0.4\text{ V} - \sqrt{(1.0\text{ V} - 0.4\text{ V})^2 - \left(\frac{0.1\text{ mA/V}^2}{2.5\text{ mA/V}^2}\right)(-2\text{ V})^2}$$

$$= 0.15\text{ V}$$

In this example, the approximate values of V_{IL} and V_{IH} are very close to the numerical results, but V_{OL} is underestimated by 40%.

12.11.2 Dissipation

Unlike the case of CMOS, the static dissipation is often dominant in DCFL. With the output at logic zero,

$$P_L = V_{DD}\beta_L(-V_{PL})^2\left[1 + \lambda_L(V_{DD} - V_{OL})\right]\tanh\left[\alpha_L(V_{DD} - V_{OL})\right]. \quad (12.31)$$

With the output at logic one, the pull-down transistor is cut off, so

$$P_H \approx 0$$

and

$$P_{DC} = \frac{1}{2}V_{DD}\beta_L(-V_{PL})^2\left[1 + \lambda_L(V_{DD} - V_{OL})\right]\tanh\left[\alpha_L(V_{DD} - V_{OL})\right]. \quad (12.32)$$

For approximate hand calculations, the static dissipation may be estimated with reasonable accuracy using

$$P_{DC} \approx \frac{1}{2}V_{DD}\beta_L(-V_{PL})^2. \quad (12.33)$$

The overall dissipation per gate is

$$P = P_{DC} + \alpha f V_{DD}^2 C_L.$$

12.11.3 Propagation Delays

The propagation delays for a DCFL gate with a load capacitance C_L may be estimated using equations analogous to those for NMOS. Thus

$$t_{PLH} \approx \frac{V_{DD}C_L}{\beta_L V_{PL}^2} \quad (12.34)$$

and

$$t_{PHL} \approx \frac{V_{DD}C_L}{\beta_O\left(V_{DD} - V_{PO}\right)^2}. \tag{12.35}$$

These approximations typically yield a factor of two accuracy. More accurate predictions can be obtained using SPICE simulations and the Curtice model. For practical DCFL circuits,

$$t_{PLH} > t_{PHL}. \tag{12.36}$$

Example 12.6

Estimate the propagation delays for the GaAs DCFL shown for the case of a 1-pF load.

Solution. The approximate delays are given by

$$t_{PLH} \approx \frac{V_{DD}C_L}{\beta_L V_{PL}^2} = \frac{(1.0\text{ V})\left(10^{-12}\text{ F}\right)}{\left(0.1\text{ mA/V}^2\right)(-2\text{ V})^2} = 2.5\text{ ns}$$

and

$$t_{PHL} \approx \frac{V_{DD}C_L}{\beta_O\left(V_{DD} - V_{PO}\right)^2} = \frac{(1.0\text{ V})\left(10^{-12}\text{ F}\right)}{\left(2.5\text{ mA/V}^2\right)(1.0\text{ V} - 0.4\text{ V})^2} = 1.11\text{ ns}.$$

Here, t_{PLH} is predicted to be more than twice t_{PHL}.

12.11.4 Logic Design

DCFL NOR gates can be constructed by placing multiple switch transistors in parallel as shown in Figure 12.15. For the NOR3 gate, the output will go low if any of the switch transistors operates in the linear mode. The output will go high if and only if all three switch transistors are cut off. DCFL NAND gates are not practical because of the tight restrictions placed on the voltages (V_{DD}, V_{PO}, V_{PL}, and V_{OL}) in these circuits.

GaAs DCFL circuits are sometimes buffered using a source follower, which results in "buffered DCFL." A buffered DCFL NOR2 gate is shown in Figure 12.16. The emitter follower M_{NP} provides a low output impedance and improves the dynamic response with highly capacitive loads. Additional pull-down transistors must also be used to allow proper function of the gate.

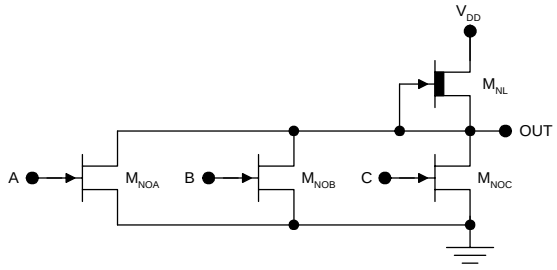


FIGURE 12.15
GaAs DCFL NOR3 gate.

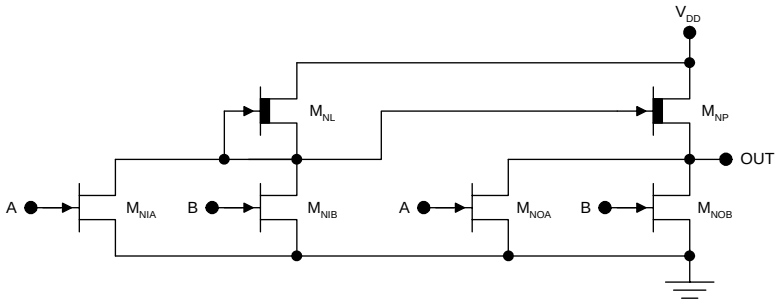


FIGURE 12.16
Buffered DCFL NOR2 gate.

12.12 PSPICE Simulations

Simulations were performed using PSPICE 9.1, which can be downloaded free.[22] The MOSFET model parameters used in the DCFL simulations are provided in Table 12.2 and Table 12.3.

TABLE 12.2

Depletion-Type GaAs MESFET[a] SPICE Parameters

Parameter	Value	Units
VTO	–2.0	V
BETA	0.1m	A/V
ALPHA	1.5	V^{-1}
LAMBDA	0.01	V^{-1}
CGS	0.6f	F
CGD	0.4f	F

[a] Pull-up device.

TABLE 12.3

Enhancement-Type GaAs[a] MESFET SPICE Parameters

Parameter	Value	Units
VTO	0.4	V
BETA	2.5m	A/V
ALPHA	1.5	V^{-1}
LAMBDA	0.01	V^{-1}
CGS	15f	F
CGD	10f	F

[a] Pull-down device.

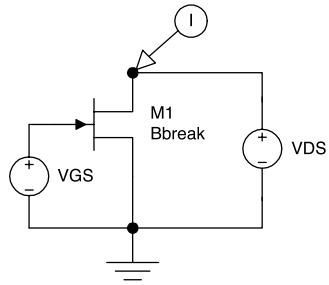


FIGURE 12.17
Circuit for the simulation of the characteristic curves of a depletion-type GaAs n-MESFET.

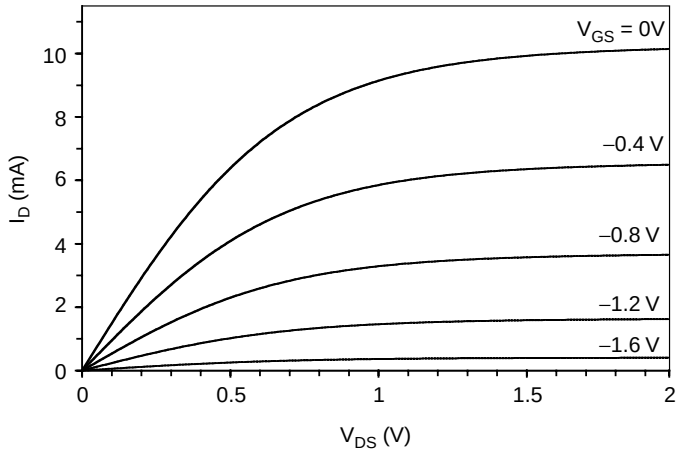


FIGURE 12.18
Simulated common source characteristics for a GaAs MESFET (I_D vs. V_{DS} with V_{GS} as a parameter).

12.12.1 GaAs MESFET Characteristics

The common source characteristics (I_D vs. V_{DS} with V_{GS} as a parameter) for a depletion-type GaAs n-MESFET were simulated using the circuit of Figure 12.17. The Bbreak part ("gasfet breakout device") was used, with VTO = –2.5, BETA = 2.5 m, ALPHA = 1.5, and LAMBDA = 0.01. The MESFET common source characteristics appear in Figure 12.18.

12.12.2 DCFL Voltage Transfer Characteristic

The voltage transfer characteristic was simulated for the DCFL inverter of Figure 12.19 with V_{DD} = 1.0 V. The results are shown in Figure 12.20.

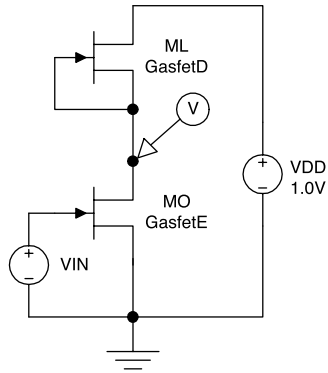


FIGURE 12.19
DCFL circuit used for the simulation of the VTC.

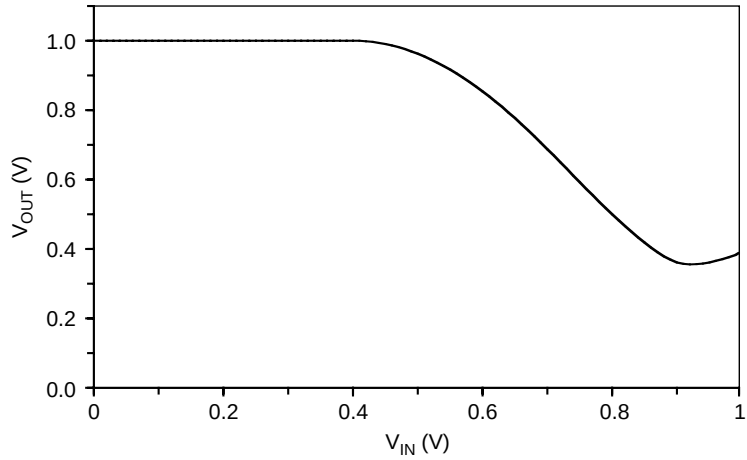


FIGURE 12.20
Simulated VTC for the DCFL circuit.

12.12.3 DCFL Propagation Delays

The propagation delays for the DCFL inverter with a 1-pF load were determined using the circuit of Figure 12.21. The pulse source parameters were V1 = 0 V, V2 = 1.0, TD = TR = TF = 0, PW = 4 ns, and PER = 10 ns.

The results of the transient simulation appear in Figure 12.22. The propagation delays can be determined using the 50% points on the input and output waveforms and are t_{PHL} = 0.50 ns and t_{PLH} = 1.45 ns. The dynamic response is asymmetric for an inverter achieving a useful value of V_{OL}. The propagation delays scale with the load capacitance. For on-chip loads, the propagation delays are typically 40 to 50% of the values achieved with CMOS using the same geometrical design rules.

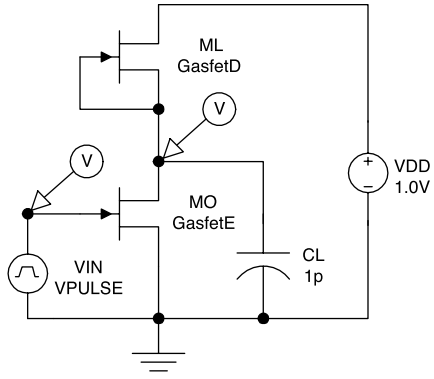


FIGURE 12.21
DCFL circuit used for simulation of the propagation delays.

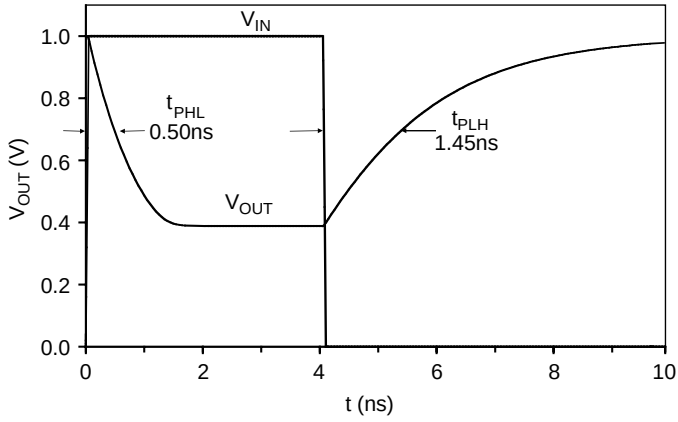


FIGURE 12.22
Simulated DCFL transient response.

12.13 Summary

GaAs digital circuits are two to three times faster than their silicon counterparts because of the superior electron mobility in GaAs. Also, GaAs gates exhibit better power delay products than those made in silicon because the peak electron velocity is achieved at lower values of the electric field intensity.

Because no viable GaAs MOS technology exists, GaAs digital integrated circuits are realized using metal–semiconductor field-effect transistors (MESFETs). The most commonly used circuit family is direct-coupled FET logic (DCFL). These logic circuits use enhancement type and depletion type MESFETs; however they run on a single supply voltage and do not require

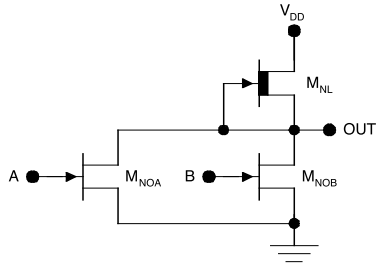
level-shifting diodes like earlier families of GaAs digital circuits. In contrast to CMOS, the DC dissipation is dominant in GaAs DCFL. Complementary circuits like CMOS are not used in GaAs, because the hole mobility in GaAs is actually inferior to that in silicon. Thus it is not possible to realize GaAs circuits that compete with silicon CMOS in terms of standby power.

An important disadvantage of GaAs DCFL is the cost. The high cost is due to the small wafer size of GaAs and relatively difficult processing and low yield compared to silicon. Additionally, lower functionality is achieved in GaAs integrated circuits in terms of logic gates per chip and embedded memory bits per chip. This is also a consequence of the difficult processing because there is a trade-off between functionality and yield.

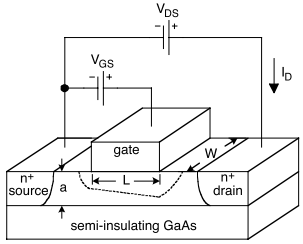
As a result of the higher cost and higher standby power compared to CMOS, GaAs DCFL applications are confined to the niches of high-end computing and high data-rate communications. This is bound to remain true as a result of the rapid pace of development in CMOS.

GaAs DIRECT-COUPLED FET LOGIC QUICK REFERENCE

NOR2 Circuit



Gallium Arsenide MESFET



GaAs circuits enjoy a three-fold speed advantage over their silicon counterparts because of the high electron mobility and peak velocity in this material. There is no viable MOS technology for GaAs so MESFETs are used. The most popular digital GaAs circuit family is DCFL.

MESFET Curtice Model

$$I_D = \begin{cases} 0; & V_{GS} < V_P \\ \beta (V_{GS} - V_P)^2 (1 + \lambda V_{DS}) \tanh(\alpha V_{DS}); & V_{GS} \geq V_P \end{cases}$$

$$V_P = V_{bi} - \frac{qN_d a^2}{2\varepsilon} \qquad \beta = \frac{\mu_n \varepsilon_s}{2a} \frac{W}{L}$$

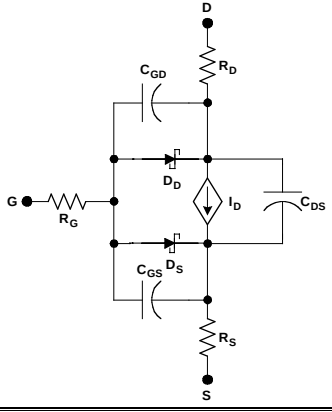
Voltage Transfer Characteristic

$$V_{IN} = V_{PO} + \sqrt{\frac{\beta_L (-V_{PL})^2 [1 + \lambda_L (V_{DD} - V_{OUT})] \tanh[\alpha_L (V_{DD} - V_{OUT})]}{\beta_O (1 + \lambda_O V_{OUT}) \tanh(\alpha_O V_{OUT})}}$$

Approximate Relationships: NMOS Analogy

$$t_{PLH} \approx \frac{V_{DD} C_L}{\beta_L V_{PL}^2} \qquad t_{PHL} \approx \frac{V_{DD} C_L}{\beta_O (V_{DD} - V_{PO})^2} \qquad P_{DC} \approx \frac{1}{2} V_{DD} \beta_L (-V_{PL})^2$$

MESFET SPICE Model



Drain Current

$$I_D = BETA(V_{GS} - VTO)^2 \times (1 + LAMBDAV_{DS}) \tanh(ALPHAV_{DS})$$

Parasitic diode currents

$$I_{DD} = IS\left[\exp\left(\frac{V_{GD}}{NV_T}\right) - 1\right]$$

$$I_{DS} = IS\left[\exp\left(\frac{V_{GS}}{NV_T}\right) - 1\right]$$

f = 10^{-15}	p = 10^{-12}	n = 10^{-9}	μ = 10^{-6}	m = 10^{-3}	k = 10^{3}	M = 10^{6}	G = 10^{9}

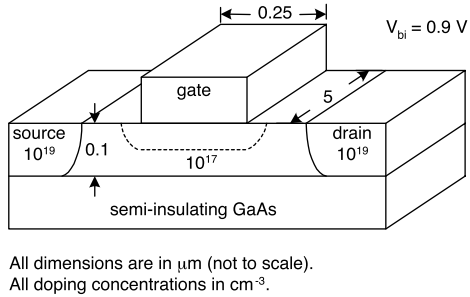


FIGURE 12.23
GaAs MESFET (P12.2).

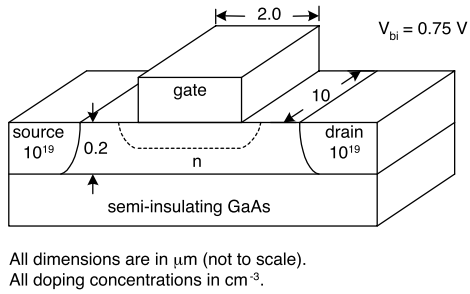


FIGURE 12.24
GaAs MESFET (P12.3).

Problems

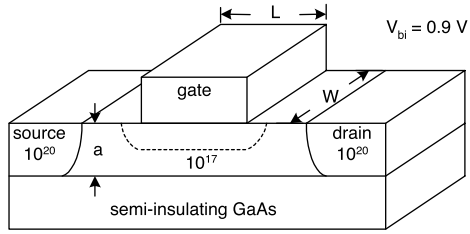
P12.1. Estimate the transit times for electrons in GaAs and Silicon FETs with 0.1 μm gate lengths. Assume that the electrons are traveling at the peak velocity in both cases.

P12.2. Estimate the pinch-off voltage for the GaAs MESFET shown in Figure 12.23 and determine if the device is an EFET or a DFET.

P12.3. For the GaAs MESFET depicted in Figure 12.24, determine the requirement on the channel doping (including tolerances) such that $V_P = 0.5\ \text{V} \pm 0.1\ \text{V}$.

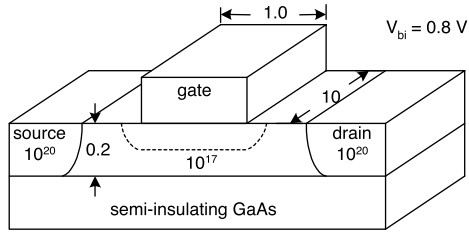
P12.4. For the GaAs MESFET illustrated in Figure 12.25, determine the requirement on the channel height (including tolerances) such that $V_P = -0.8\ \text{V} \pm 0.1\ \text{V}$.

P12.5. For the GaAs MESFET shown in Figure 12.26, estimate the tolerance in V_P if the tolerances for the channel doping and height are each ±5%.



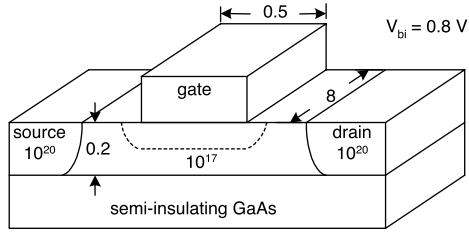
All dimensions are in μm (not to scale).
All doping concentrations in cm^{-3}.

FIGURE 12.25
GaAs MESFET (P12.4).



All dimensions are in μm (not to scale).
All doping concentrations in cm^{-3}.

FIGURE 12.26
GaAs MESFET (P12.5).



All dimensions are in μm (not to scale).
All doping concentrations in cm^{-3}.

FIGURE 12.27
GaAs MESFET (P12.6).

P12.6. For the GaAs MESFET with the design shown in Figure 12.27, estimate the device transconductance parameter.

P12.7. Calculate and plot the common source characteristics for a GaAs MESFET with $V_P = -1$ V, $\beta = 5$ mA/V^2, $\lambda = 0.01$, and $\alpha = 5$ V^{-1}.

P12.8. Using numerical calculations, determine the critical voltages for the DCFL inverter depicted in Figure 12.28.

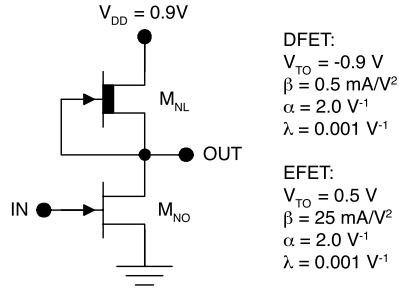


FIGURE 12.28
DCFL inverter (P12.7).

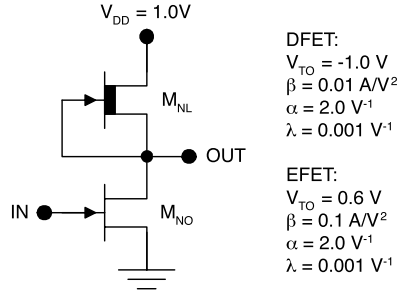


FIGURE 12.29
DCFL inverter (P12.8).

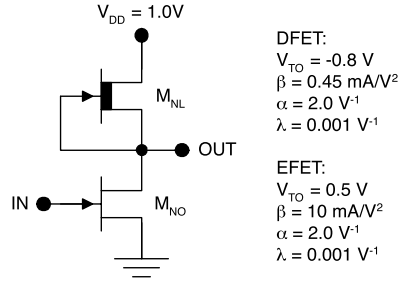


FIGURE 12.30
DCFL inverter (P12.9).

P12.9. Determine and plot the voltage transfer characteristic for the DCFL inverter of Figure 12.29 and determine V_{OL}.

P12.10. Consider the DCFL inverter illustrated in Figure 12.30.

1. Estimate the critical voltages using the NMOS analogy.
2. Determine the voltage transfer characteristic using the Curtice model for both transistors. Determine the critical voltages and compare the values to the estimates obtained using the NMOS analogy.

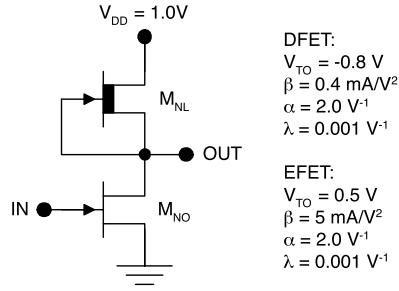


FIGURE 12.31
DCFL inverter (P12.10).

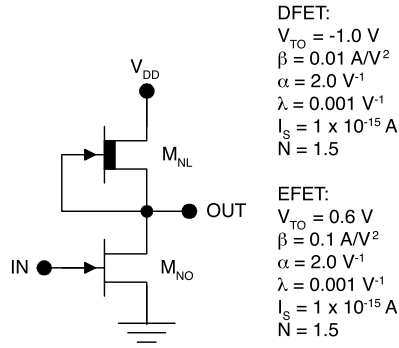


FIGURE 12.32
DCFL inverter (P12.12).

P12.11. For the DCFL inverter shown in Figure 12.31, estimate the average DC dissipation using the NMOS analogy. Also determine the average DC dissipation using the Curtice model and numerical calculations; compare the two results.

P12.12. For the type of DCFL inverter shown in Figure 12.32, determine the maximum supply voltage that can be used if the DC gate current should not exceed 1 μA.

References

1. Deyhimy, I., Gallium arsenide joins the giants, *IEEE Spectrum*, 32, 33, Feb. 1995.
2. Brown, R.B., Barker, P., Chandna, A., Huff, T.R., Kayssi, A.I., Lomax, R.J., Mudge, T.N., Nagle, D., Sakallah, K.A., Sherhart, P.J., Uhlig, R., and Upton, M., GaAs RISC processors, *Tech. Dig. 1992 GaAs IC Symp.*, 81, 1992.
3. Harrington, D.L., Troeger, G.L., Gee, W.C., Bolen, J.A., Vogelsang, C.H., Nicalek, T.P., Lowe, C.M., Roh, Y.K., Nguyen, K.Q., Fay, J.F., and Reeder, J., A GaAs 32-bit RISC microprocessor, *Tech. Dig. 10th GaAs IC Symp.*, 87, 1988.

4. Onodera, T., Onodera, H., Sugisaki, S., Okamoto, M., Suyama, K., Kuryu, I., and Nishi, H., GaAs MESFET LSI design using E/D-DCFL circuits, *Tech. Dig. 12th GaAs IC Symp.*, 219, 1990.
5. Mudge, T.N., Brown, R.B., Birmingham, W.P., Dykstra, J.A., Kayssi, A.I., Lomax, R.J., Olukotun, O.A., Sakallah, K.A., The design of a GaAs micro-supercomputer, system sciences, *Proc. 24th Annu. Hawaii Int. Conf. Sys. Sci.*, 1, 421, 1991.
6. Yanyang, X., Xiaoguang, Z., and Jingchen, H., Direct coupled FET logic (DCFL) circuit for GaAs LSIC application, *Proc. 1998 Int. Conf. Microwave Millimeter Wave Technol.*, 913, 1998.
7. Shikata, M., Tanaka, K., Yamada, H.T., Fujishiro, H.I., Nishi, S., Yamagishi, C., and Akiyama, M., A 20-Gb/s flip-flop circuit using direct-coupled FET logic, *IEEE J. Solid-State Circuits*, 28, 1046, 1993.
8. Vu, T.T., Nelson, R.D., Lee, G.M., Roberts, P.C.T., Lee, K.W., Swanson, S.K., Peczalski, A., Betten, W.R., Hanka, S.A., Helix, M.J., Vold, P.J., Lee, G.Y., Jamison, S.A., Arsenault, C.A., Karwoski, S.M., Naused, B.A., Gilbert, B.K., and Shur, M.S., Low-power 2K-cell SDFL gate array and DCFL circuits using GaAs self-aligned E/D MESFETs, *IEEE J. Solid-State Circuits*, 23, 224, 1988.
9. Kayssi, A.I. and Sakallah, K.A., Delay macromodels for the timing analysis of GaAs DCFL, *1992 Design Automation Conf.*, 142, 1992.
10. Higashisaka, N., Shimada, M., Ohta, A., Hosogi, K., Tobita, Y., and Mitsui, Y., GaAs DCFL 2.5 Gbps 16-bit multiplexer/demultiplexer LSIs, *IEEE J. Solid-State Circuits*, 29, 808, 1994.
11. Garcia, J., Hernandez, A., del Pino, J., Sendra, J.R., Gonzalez, B., and Nunez, A., Power model for DCFL family, *Electron. Lett.*, 38, 13, 2002.
12. Nemoto, M., Ogawa, Y., Morita, Y., Seki, S., Kawakami, Y., and Akiyama, M., 25 ps/gate GaAs standard cell LSIs using 0.5 μm gate MESFETs, *Tech. Dig. 14th GaAs IC Symp.*, 93, 1992.
13. Hinds, R., Canaga, S., Lee, G., and Choudhury, A., A 20 K GaAs array with 10 K of embedded SRAM, *Proc. 1990 IEEE Custom Integrated Circuits Conf.*, 1, 1990.
14. Mikkelson, J., GaAs digital VLSI device and circuit technology, *Tech. Dig. Int. Electron Devices Meet.*, 231, 1991.
15. www.vitesse.com (Vitesse Semiconductor).
16. www.lucent.com (Lucent Technologies).
17. www.tqs.com (TriQuint Semiconductor).
18. www.alphaind.com (Alpha Industries).
19. Sze, S.M., *Physics of Semiconductor Devices*, John Wiley & Sons, New York, 1981.
20. Curtice, W.R., A MESFET model for use in the design of GaAs integrated circuits, *IEEE Trans. Microwave Theory Tech.*, 28, 448, 1980.
21. Statz, H., Newman, P., Smith, I.W., Pucel, R.A., and Haus, H.A., GaAs FET device and circuit simulation in SPICE, *IEEE Trans. Electron. Devices*, 34, 160, 1987.
22. www.cadence.com (Cadence).

13

Interfacing between Digital Logic Circuits

13.1 Introduction

Thus far the emphasis of this book has been on the design and analysis of gate circuits. However, the implementation of digital systems involves a number of other circuit considerations. First, most digital systems involve a mixture of circuits from different logic families. This necessitates interfacing between the different types of circuits with different current requirements and voltage levels by using level-shifting circuits.[1–13] Second, it is often necessary to connect the outputs of multiple gates to a single node, as is the case with data, address, and control busses, and other signal lines. In some cases, all such outputs can remain active and the resulting level on the line is determined by the ORing or ANDing of the multiple outputs, which is referred to as *wired logic*.[14] In other cases, it is necessary to allow only one of the outputs to be active at a particular time; this is achieved using *transmission gates*[15] or *tri-state logic gates*.[16–22]

13.2 Level-Shifting Circuits

Many digital systems involve a mixture of circuits from different circuit families. This choice is made by the designer in order to optimize different subsystems with different requirements simultaneously. For example, if a CMOS processor may be connected to TTL bus drivers and NMOS memory circuits, the processor utilizes CMOS for high packing density, low power, and short on-chip propagation delays. The TTL bus drivers are used for high off-chip data rates and the NMOS memory circuits are used for maximum density and number of bits. In practice, because all conceivable circuit combinations have been used, interfacing the different types of circuits with different current requirements and voltage levels is necessary.

The two general requirements of an interface circuit are level shifting and buffering. The first requirement stems from the fact that different circuit

families use different voltage levels to represent "1" and "0." The buffering requirement refers to the need to provide sufficient drive current for the destination circuit. In meeting these two general requirements, it is necessary to give due consideration to the speed of the interface circuit. A slow interface circuit could create a "bottleneck," thus adversely affecting the rate of data flowing between the subsystems.

Numerous interfacing circuits are in use and no attempt will be made to catalog them here. Instead, a few examples will be discussed to illustrate the general principles outlined.

13.2.1 ECL to TTL

As a first example, consider the interface from ECL to TTL. Here the primary requirement is a translation from the negative voltage levels of ECL to the positive levels used by TTL. For this reason, the interface circuit is sometimes called an "ECL-to-TTL level translator" (Figure 13.1). For this ECL-to-TTL level translator, if $V_{IN} = V_{OL}$ (ECL), then Q_I is cut off. Therefore, current flows to the base of Q_O, turning it "on hard" and bringing the output to $V_{BEOH} = V_{OL}$ (TTL). On the other hand, if $V_{IN} = V_{OH}$ (ECL), then Q_I turns on and brings the base voltage of Q_O sufficiently low so that that transistor is cut off. As a result, the output rises to V_{DD}, which is interpreted as logic one by TTL circuits. The basic operation of the circuit is summarized in Table 13.1.

Circuit speed is critical in the ECL-to-TTL level translator. A sluggish translator could negate the benefits of using ECL in the first place, so Q_O has been

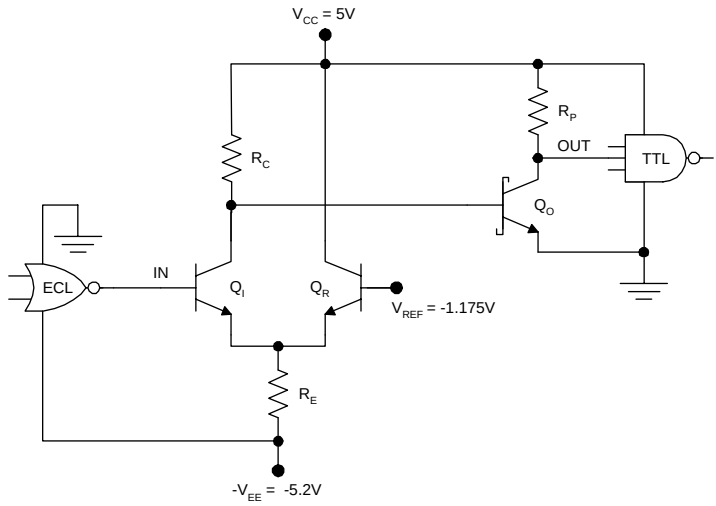


FIGURE 13.1
ECL-to-TTL level translator.

TABLE 13.1

Basic Operation of the ECL-to-TTL Level Translator

V_{IN}	Q_I	Q_R	Q_O	V_{OUT}
V_{OL}(ECL)	CO[a]	FA	OH[c]	V_{CEOH}
V_{OH}(ECL)	FA[b]	CO	CO	V_{CC}

[a] CO = cutoff.
[b] FA = forward active.
[c] OH = on hard.

TABLE 13.2

Basic Operation of the TTL-to-ECL Level Translator

V_{IN}	Q_I	Q_R	Q_{INV}	V_{OUT}
V_{OL}(TTL)	CO[a]	FA	FA	V_{OL}(ECL)
V_{OH}(TTL)	FA[b]	CO	FA	V_{OH}(ECL)

[a] CO = cutoff.
[b] FA = forward active.

Schottky clamped. In addition, this simple circuit must be used in conjunction with a TTL buffer to minimize the capacitive loading at the output. With the passive pull-up scheme shown, care must be taken to keep the R_PC_L product less than the propagation delay for the ECL.

13.2.2 TTL to ECL

As with the last example, because level translation is the most important requirement for a TTL-to-ECL interface, such an interface is sometimes called a TTL-to-ECL level translator. Such a translator is shown in Figure 13.2; the basic operation is as follows. If $V_{IN} = V_{OL}$ (TTL), then the voltage at the base of Q_I drops below V_{REF} so that Q_I is cut off. Q_R conducts, so the output goes low as in a normal ECL gate. On the other hand, if $V_{IN} = V_{OH}$ (TTL), then Q_I turns on, causing Q_R to cut off, and the output goes high as in a normal ECL gate. The basic operation of this circuit is summarized in Table 13.2.

13.2.3 High-Voltage CMOS to Low-Voltage CMOS

Large digital systems typically use CMOS circuits running at different supply voltages. This is true even within a single chip where a processor core might run at a low supply voltage to minimize dissipation but the output drivers may run at a much higher voltage to provide the required off-chip data rates.

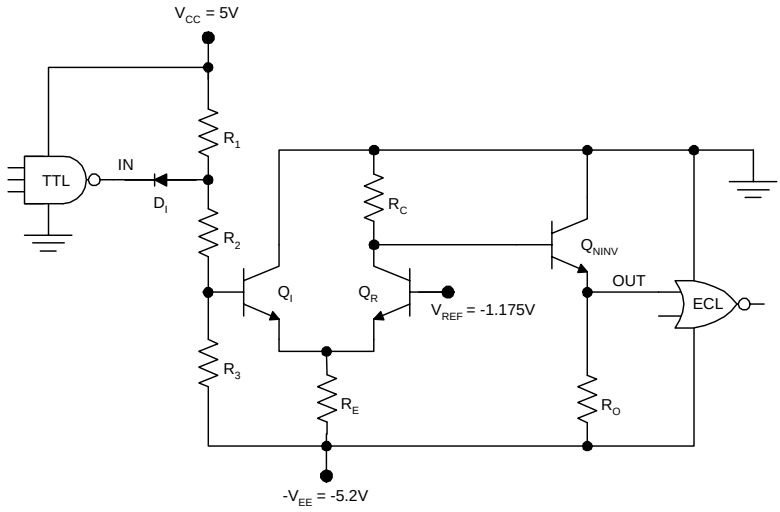


FIGURE 13.2
TTL-to-ECL level translator.

Interfacing from high-voltage CMOS (HV CMOS) to low-voltage CMOS (LV CMOS) requires a level translator. One way to achieve such a level translator is by an asymmetric CMOS inverter. In a symmetric CMOS inverter ($K_N = K_p = K$ and $V_{TN} = |V_{TP}| = V_T$), the midpoint in the VTC as defined by $V_{IN} = V_{OUT} = V_M$ is

$$V_M = V_{DD}/2 . \tag{13.1}$$

However, in an asymmetric CMOS inverter with $K_N \neq K_p$ but $V_{TN} = |V_{TP}| = V_T$, the midpoint is given by

$$V_M = \frac{V_T + (V_{DD} - V_T)\sqrt{K_P/K_N}}{1 + \sqrt{K_P/K_N}} . \tag{13.2}$$

Rearranging, the ratio of the device transconductance parameters is:

$$K_R = \frac{K_N}{K_P} = \left(\frac{V_{DD} - V_T - V_M}{V_M - V_T} \right)^2 . \tag{13.3}$$

A CMOS inverter interfacing from HV CMOS to LV CMOS can be designed using the lower supply voltage V_{DDL} so that V_{OH} will be equal to supply voltage of the gate being driven. K_R is chosen so that V_M is equal to one half of the higher supply voltage V_{DDH}. Then the required ratio of device transconductance parameters is

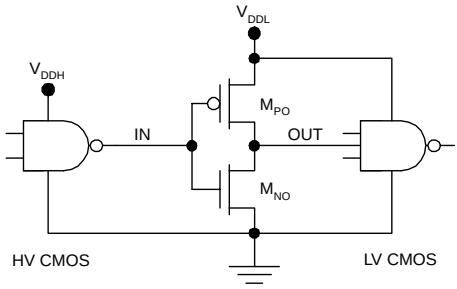


FIGURE 13.3
HV CMOS-to-LV CMOS level shifter.

$$K_R = \left(\frac{V_{DDL} - V_{DDH}/2 - V_T}{V_{DDH}/2 - V_T} \right)^2, \tag{13.4}$$

which results in the simple interface scheme shown in Figure 13.3.

A limitation of this approach is that it is not possible to match the midpoint of the VTC to V_{DDH} if $V_{DDH}/2 > V_{DDL} - V_T$. In such cases, more than one stage may used if multiple supply voltages are available.

Example 13.1

Design a CMOS level shifter to interface from 5-V circuits to 3.3-V circuits and plot the VTC for the circuit. Assume that V_T = 0.3 V for all circuits.

Solution. To interface from 5-V circuits to 3.3-V circuits, use $V_{DD} = V_{DDL}$ = 3.3 V and

$$K_R = \left(\frac{V_{DDL} - V_{DDH}/2 - V_T}{V_{DDH}/2 - V_T} \right)^2 = \left(\frac{3.3\text{ V} - 5\text{ V}/2 - 0.3\text{ V}}{5\text{ V}/2 - 0.3\text{ V}} \right) = 0.051\,.$$

Therefore, $K_P/K_N \approx 19$.

The voltage transfer characteristic can be calculated piecewise as

$$V_{\text{OUT}} = \begin{cases} 3.3\text{ V}; & V_{\text{IN}} \le 0.3\text{ V} \\ V_{\text{IN}} + 0.3\text{ V} + \sqrt{\left(V_{\text{IN}} - 3.0\text{ V}\right)^2 - (1/19)\left(V_{\text{IN}} - 0.3\text{ V}\right)^2}\,; & 0.3\text{ V} \le V_{\text{IN}} \le 2.5\text{ V} \\ 1.65\text{ V}; & V_{\text{IN}} = 2.5\text{ V} \\ V_{\text{IN}} - 0.3\text{ V} - \sqrt{\left(V_{\text{IN}} - 0.3\text{ V}\right)^2 - (19)\left(V_{\text{IN}} - 3.0\text{ V}\right)^2}\,; & 2.5\text{ V} \le V_{\text{IN}} \le 4.7\text{ V} \\ 0; & V_{\text{IN}} \ge 4.7\text{ V} \end{cases}$$

The voltage transfer characteristic appears in Figure 13.4. V_M (and therefore V_{IL} and V_{IH}) corresponds approximately to the values for symmetric 5-V CMOS circuits. On the other hand, the output levels V_{OL} and V_{OH} correspond to 3.3-V CMOS circuits.

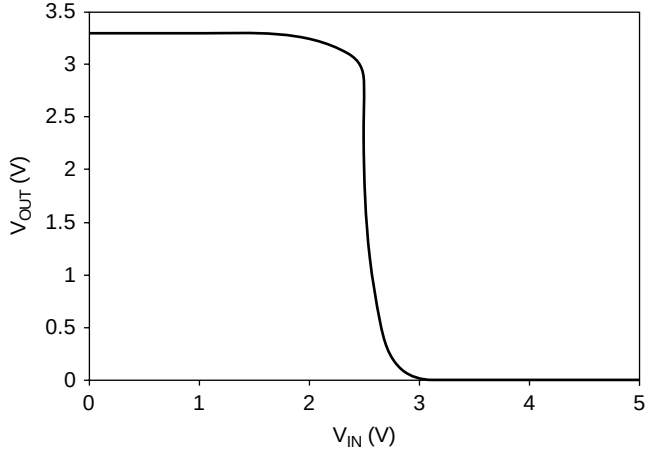


FIGURE 13.4
Voltage transfer characteristic for a 5-V CMOS to 3.3-V CMOS level translator.

13.2.4 Low-Voltage CMOS to High-Voltage CMOS

Level shifting from LV CMOS circuits to HV CMOS circuits can also be accomplished using an asymmetric CMOS inverter. Here the level shifter must operate at the higher supply voltage so that its value of V_{OH} matches the driven gate. A problem with this approach is that the logic-one level from the LV CMOS circuits is insufficient to cut off the p-MOSFET in the level translator. If the two supply voltages are very different, this leads to appreciable static conduction in the level translator.

Another approach is to use a pseudo NMOS inverter cascaded with a CMOS inverter as shown in Figure 13.5. This design reduces, but does not eliminate, the static conduction because static conduction is inherent in pseudo NMOS gates. However, the pseudo NMOS gate drives only the CMOS inverter, so it can be scaled down considerably to minimize the static dissipation.

Following this approach, the pseudo NMOS stage should be designed using $V_{DD} = V_{DDH}$ and $V_M = V_{DDL}/2$. This requires that

$$\frac{K_{NI}}{K_{PI}} = \frac{\left(V_{DDH} - V_T\right)\left(V_{DDH} - V_{DDL}/2\right) - \left(V_{DDH} - V_M\right)^2/2}{\left(V_{DDL}/2 - V_T\right)^2/2}. \tag{13.5}$$

The absolute values of the device transconductance parameters in the pseudo NMOS stage must be chosen so that the CMOS stage can be driven at the necessary bitrate. This stage can be designed with symmetric transistors, but should be scaled based on the load capacitance as well as the bitrate requirement. The same approach can be used to interface from LV CMOS to HV BiCMOS drivers.

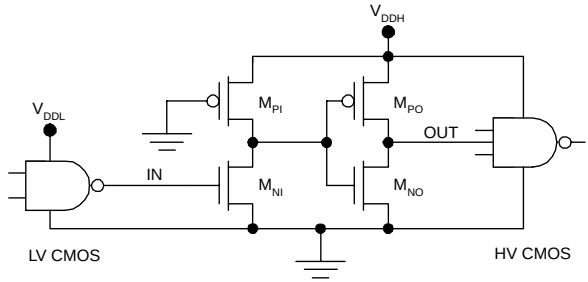


FIGURE 13.5
LV CMOS-to-HV CMOS level translator utilizing a pseudo NMOS inverter driving a CMOS inverter.

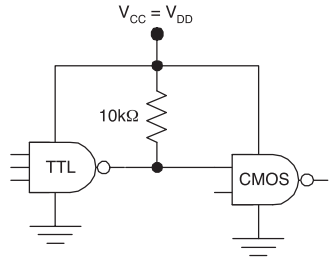


FIGURE 13.6
Interfacing TTL to CMOS.

13.2.5 TTL to CMOS

Interfacing TTL to CMOS is straightforward if the two circuit families use the same supply voltage. In this case, the voltage levels are almost identical. However, because V_{OH} (TTL) is less than $V_{DD} - V_T$ for the CMOS, a pull-up resistor is necessary at the interface, as shown in Figure 13.6. This pull-up resistor allows the interface node to rise to V_{DD} with a high output from the-TTL. This allows the p-MOSFET in the CMOS gate to turn off fully under this condition. Current buffering is not a problem because of the low current draw of CMOS.

If TTL gates drive LV CMOS circuits, it is often possible to connect without any special interface circuit. However, it may be desirable to use an asymmetric CMOS inverter as in the case of the HV CMOS-to-LV CMOS interface.

13.2.6 CMOS to TTL

The CMOS-to-TTL interface is almost trivial if the two circuit families use the same supply voltage; because CMOS has rail-to-rail voltage swing, no level translation is necessary. Current buffering is occasionally needed due to the DC loading presented by TTL. Otherwise, the two circuit families may

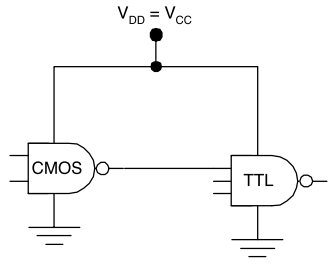


FIGURE 13.7
Interfacing TTL to CMOS.

be connected directly as shown in Figure 13.7. Sometimes it is necessary to interface between LV CMOS and TTL circuits. The methods used in LV CMOS-to-HV CMOS interfaces may be employed in such cases.

13.3 Wired Logic

In some situations, it is possible to connect the outputs of two or more logic gates, creating a new logic function without adding a physical logic circuit. This is referred to as wired logic. Because wired logic allows the addition of functionality without the addition of gates, the potential benefits include reduced cost, package count, size, weight, and dissipation.

However, wired logic is not possible with gates that have active pull-up and active pull-down, e.g., CMOS. If the outputs of two CMOS gates are connected as shown in Figure 13.8 and one gate tries to drive the output node high while the other tries to drive the output low, a disastrous short circuit will result. The ensuing high current is likely to cause irreversible damage to the output transistors. The same is true of any logic family with active pull-up and pull-down, including TTL.

On the other hand, wired logic can be used in the case of "open collector"-TTL, a type of TTL that lacks active pull-up circuitry (Figure 13.9). If the outputs of two or more such gates are connected, the common output node will go low if any one of the individual pull-down transistors turns on.

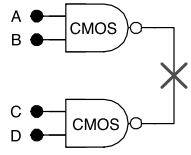


FIGURE 13.8
Wired logic is not possible with CMOS, which has active pull-up and active pull-down.

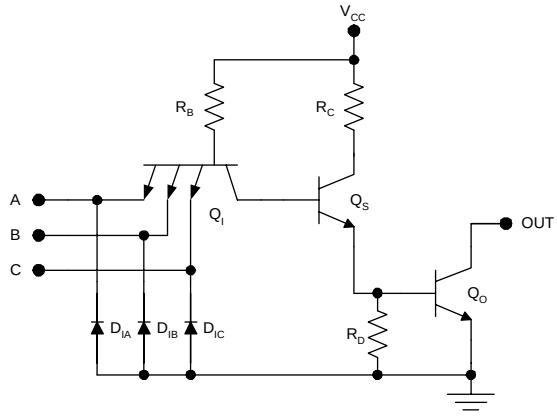


FIGURE 13.9
Open collector TTL NAND3 gate.

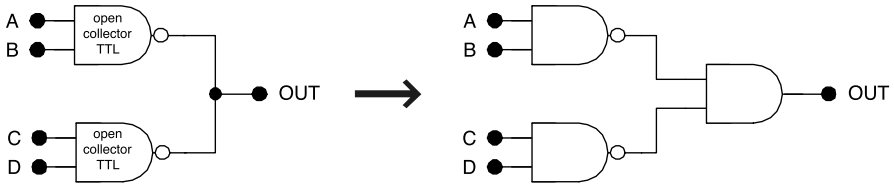


FIGURE 13.10
Wired logic using open collector TTL creates the AND function.

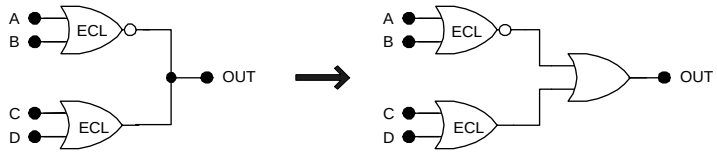


FIGURE 13.11
Wired logic using ECL gates creates the OR function.

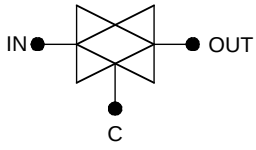
Therefore, wiring these outputs together creates the AND function shown in Figure 13.10. Often ECL gates have active pull-up but not active pull-down; therefore, wired logic is possible with these circuits, although the created function is OR rather than AND. This is because the common output node will go high if any one of the individual gates pulls the output high. This is shown in Figure 13.11.

To summarize, wired logic creates the AND function in gates with active pull-down and the OR function in circuits with active pull-up. Wired logic is not possible in circuits with active pull-up and active pull-down (Table 13.3).

TABLE 13.3

Wired Logic Summary

Output Circuitry	Wired Logic Function
Active pull-down	AND
Active pull-up	OR
Active pull-up and active pull-down	Wired logic cannot be used



C	IN	OUT
0	X	HIGH Z
1	0	0
1	1	1
X = "don't care"		

FIGURE 13.12
Transmission gate.

13.4 Transmission Gates

Sometimes it is necessary to actively disconnect the output of a logic gate from a common node and allow another gate to drive the node, which can be achieved using a transmission gate. Doing so prevents driving conflicts such as those described in the previous section.

In general, a transmission gate has a signal input, a control input, and an output, as shown in Figure 13.12. The basic operation is as follows. If logic one is applied at the control input, then the output follows the signal input. If logic zero is applied at the control input, then the output goes into a high-impedance state, whereby the signal input is essentially disconnected from the output. This function is therefore useful in data switches and multiplexers.

A CMOS transmission gate can be implemented using the circuit of Figure 13.13. The basic operation is as follows. If $C = 0$, then M_{NO} and M_{PO} are cut off. The gate operates in the high Z state. If $C = V_{DD}$, M_{NO} and M_{PO} are linear and the output follows the signal input. This operation is summarized in Table 13.4. In some cases, complementary control inputs are available, for example, in low-power adiabatic CMOS circuits. Then the inverter may be omitted and only two MOSFETs are required to make the transmission gate.

13.5 Tri-State Logic

Many digital systems route signals on busses, including data, address, and control busses. Figure 13.14 shows the example of a microprocessor system

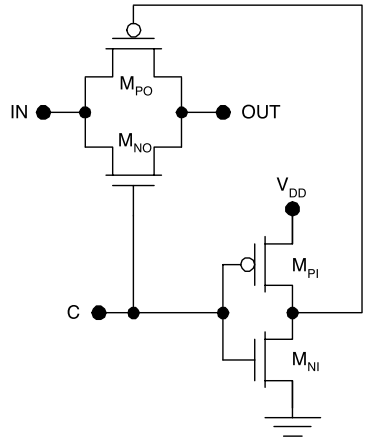


FIGURE 13.13
CMOS transmission gate.

TABLE 13.4

Basic Operation of the CMOS Transmission Gate

C	IN	M_{PO}	M_{NO}	OUT
0	0	CO[a]	CO	HIGH Z
0	V_{DD}	CO	CO	HIGH Z
V_{DD}	0	LIN[b]	LIN	0
V_{DD}	V_{DD}	LIN	LIN	V_{DD}

[a] CO = cutoff.
[b] LIN = linear.

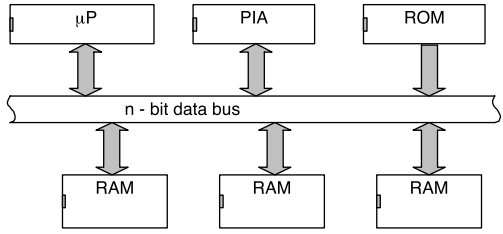


FIGURE 13.14
Bus-based system with tri-state devices.

in which multiple driving devices share the common copper lines of the bus. If conflicts are to be avoided, only one of the devices may attempt to drive the bus at a particular time. The practical solution to this problem is to use tri-state logic gates.

Tri-state logic gates exhibit three output states: 0, 1, and high Z. In the high Z, or "high impedance" state, the output is disabled. Therefore, the output

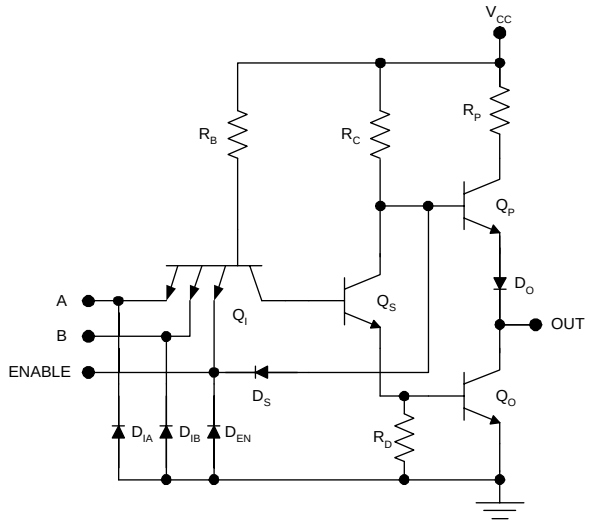


FIGURE 13.15
Tri-state TTL NAND2 gate.

node may float to whatever voltage is set by the enabled output of another device on the bus. These gates can be implemented in two ways: 1) cascade a two-state gate with a transmission gate or 2) make internal modifications to the two-state gate. Although the first approach is more straightforward, the latter is preferred because it can result in simpler, faster circuits.

A TTL gate may be modified for tri-state operation by adding a diode and one extra emitter in the input transistor. This is illustrated in Figure 13.15 for a standard TTL NAND2 gate. In this tri-state TTL gate, the application of logic zero at the enable input does two things. First, it saturates the input transistor, therefore ensuring that Q_S and Q_O will be cut off. Second, it brings the base of Q_P down to about 0.7 V, ensuring that the pull-up transistor is cut off. Therefore, application of logic zero to the enable input causes the pull-down and pull-up transistors to cut off, putting the output in the high Z state. On the other hand, application of logic one to the enable input reverse biases the base–emitter junction of Q_I and the diode D_S so that normal operation of the gate is possible. This function is summarized in Table 13.5.

Although the previous illustration used a standard TTL circuit for simplicity, the same design concepts may be applied to other TTL circuit types. The only real difference is that the other designs use an extra input diode in place of the extra emitter in Q_I.

Tri-state function can be provided in any CMOS gate by adding four MOSFETs, as illustrated in Figure 13.16 for the case of the inverter. Here, M_{NE} and M_{PE} make it possible to disconnect the core gate (comprising M_{NO} and M_{PO}) from the rails. Thus, if the enable is low, M_{NE} and M_{PE} are cut off,

TABLE 13.5

Basic Operation of the Tri-State TTL NAND2 Gate

ENABLE	A	B	Q_I	Q_S	Q_O	Q_P	OUT
0	0	0	SAT[a]	CO[b]	CO	CO	HIGH Z
0	0	V_{OH}	SAT	CO	CO	CO	HIGH Z
0	V_{OH}	0	SAT	CO	CO	CO	HIGH Z
0	V_{OH}	V_{OH}	SAT	CO	CO	CO	HIGH Z
V_{OH}	0	0	SAT	CO	CO	FA[c]	V_{OH}
V_{OH}	0	V_{OH}	SAT	CO	CO	FA	V_{OH}
V_{OH}	V_{OH}	0	SAT	CO	CO	FA	V_{OH}
V_{OH}	V_{OH}	V_{OH}	RA[d]	SAT	SAT	CO	~0

[a] SAT = saturated.
[b] CO = cutoff.
[c] FA = forward active.
[d] RA = reverse active.

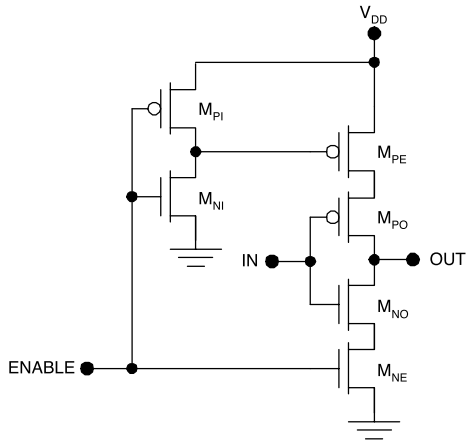


FIGURE 13.16
Tri-state CMOS inverter.

providing high Z operation. If the enable is high, M_{NE} and M_{PE} are linear, allowing normal operation of the core gate. The same basic design can be used for other CMOS gates, such as the NAND2 circuit shown in Figure 13.17. Here, too, only four extra MOSFETs are required to provide the tri-state function. Although this concept can be extended to CMOS gates with any level of complexity, due consideration must be given to scaling of the MOSFETs. The same design rules discussed in Chapter 9 apply here as well.

In CMOS buffers/output drivers, the MOSFETs are so wide that the addition of two power switching transistors consumes considerable silicon area. For this reason, tri-state CMOS buffers are usually realized as shown in Figure 13.18.

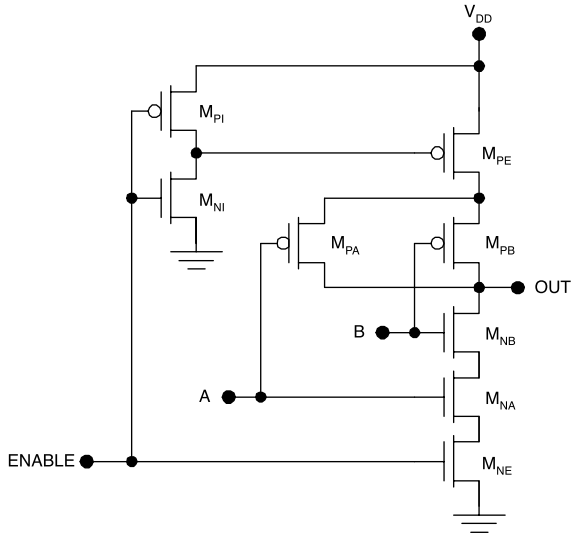


FIGURE 13.17
Tri-state CMOS NAND2 gate.

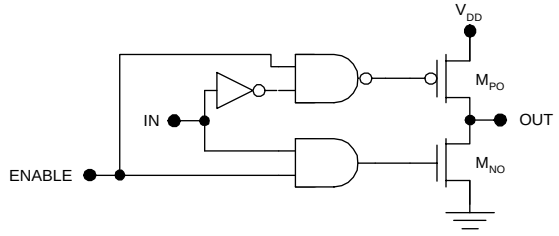


FIGURE 13.18
Tri-state CMOS driver.

13.6 PSPICE Simulations

Simulations were performed using PSPICE 9.1, which can be downloaded free.[23] The device model parameters used in all simulations are provided in Table 13.6 through Table 13.9.

13.6.1 ECL-to-TTL Level Translator

The voltage transfer characteristic for an ECL-to-TTL level translator was simulated using the circuit of Figure 13.19. The results of the DC sweep shown in Figure 13.20 demonstrate that the circuit translates from ECL

TABLE 13.6

BJT SPICE Parameters

Parameter	Value	Units
IS	2.0f	A
BF	70	—
NF	1	—
BR	0.5	—
NR	1	—
CJE	0.3p	F
VJE	0.8	V
MJE	0.333	—
TF	0.2n	s
CJC	0.15p	F
VJC	0.75	V
MJC	0.5	—
TR	10n	s

TABLE 13.7

Schottky Diode SPICE Parameters

Parameter	Value	Units
IS	9.0n	A
N	1	—
TT	0	s
CJO	0.05p	F
VJ	0.4	V
M	0.5	—

TABLE 13.8

n-MOSFET SPICE Parameters

Parameter	Value	Units
VTO	0.5	V
KP	0.2m	A/V^2
LAMBDA	0.05	—
CGSO	1.15n	F/m
CGDO	0.58n	F/m
VMAX	100k	m/s

TABLE 13.9

p-MOSFET SPICE Parameters

Parameter	Value	Units
VTO	–0.5	V
KP	0.2m	A/V^2
LAMBDA	0.05	—
CGSO	1.15n	F/m
CGDO	0.58n	F/m
VMAX	80k	m/s

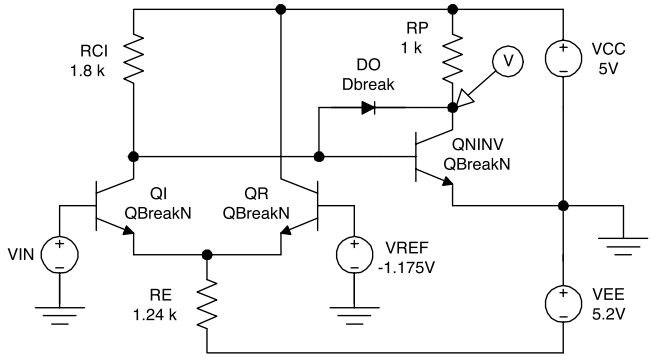


FIGURE 13.19
ECL-to-TTL level translator circuit used for simulation of the voltage transfer characteristic.

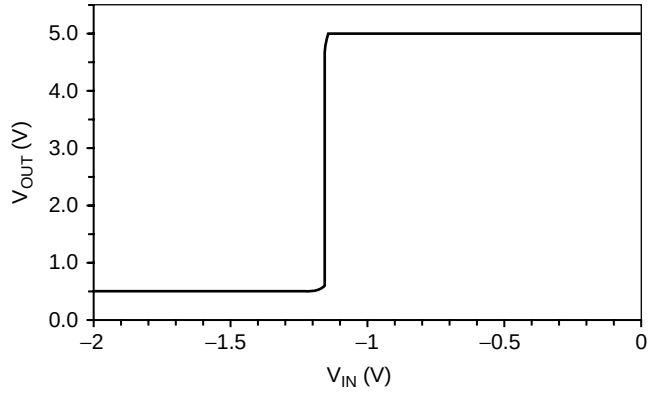


FIGURE 13.20
Simulated VTC for the ECL-to-TTL level translator.

voltage levels to TTL levels as intended. The critical voltages are V_{IL} = –1.13 V, V_{IH} = –1.11 V, V_{OL} = 0.50 V and V_{OH} = 5.00 V.

13.6.2 TTL-to-ECL Level Translator

The voltage transfer characteristic for an ECL-to-TTL level translator was simulated using the circuit of Figure 13.21. The results of the DC sweep shown in Figure 13.22 demonstrate that the circuit translates from TTL voltage levels to ECL levels as intended. V_{IL} = 2.39 V, V_{IH} = 2.78 V, V_{OL} = –1.56 V and V_{OH} = –0.98 V.

13.6.3 Tri-State TTL Inverter

The behavior of a tri-state TTL inverter was simulated using the circuit of Figure 13.23. Bipolar transistors with multiple emitters are not available in

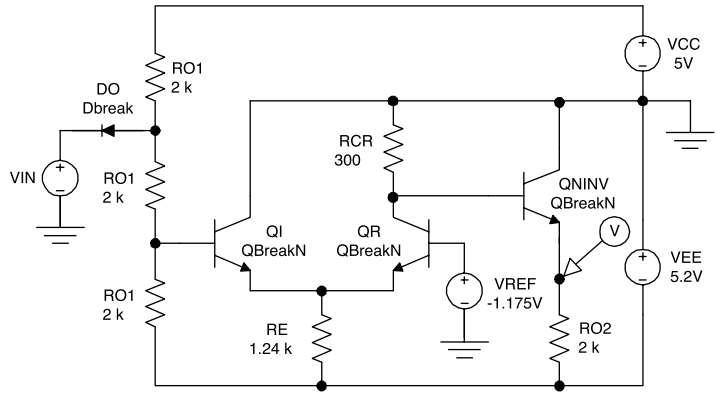


FIGURE 13.21
TTL-to-ECL level translator circuit used for simulation of the voltage transfer characteristic.

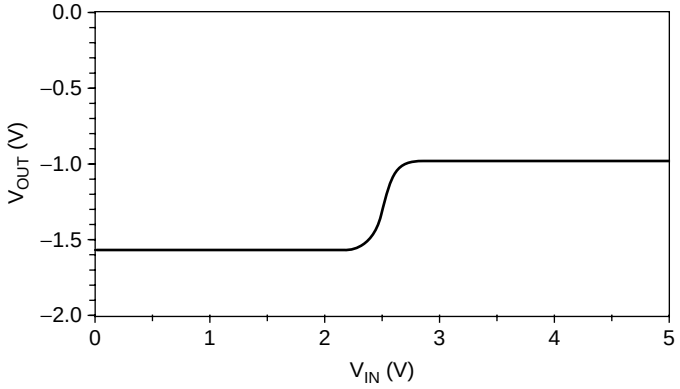


FIGURE 13.22
Simulated VTC for the TTL-to-ECL level translator.

SPICE, so the two-emitter bipolar transistor is modeled in SPICE using two single-emitter transistors connected in parallel. Pulse sources were used for the input and the enable voltages and the input was set to 10 times the frequency of the enable. The results of the transient simulation are provided in Figure 13.24. With logic one applied at the enable, normal inverter operation is obtained; with logic zero applied at the enable, the high Z state is obtained at the output. A resistor divider network was added to the output circuitry to show this clearly. Therefore, with the totem pole output floating, the voltage at the output node goes to $V_{CC}/2$.

13.6.4 Tri-State CMOS Inverter

The circuit of Figure 13.25 was used to study the behavior of a tri-state CMOS inverter. Pulse sources were used for the input and the enable voltages and

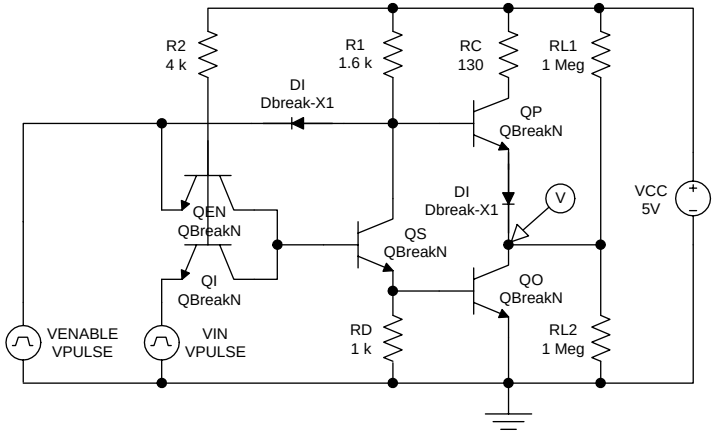


FIGURE 13.23
Tri-state TTL inverter circuit.

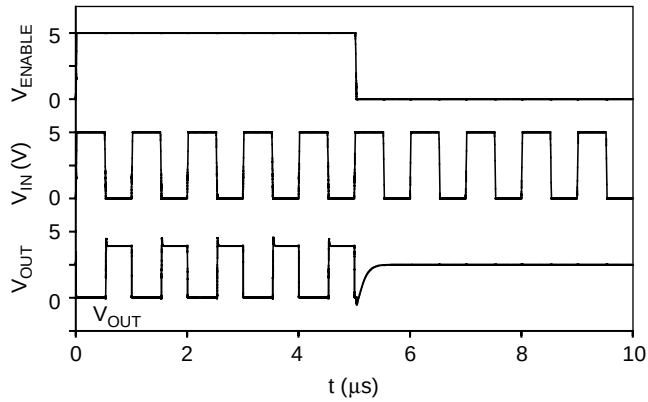


FIGURE 13.24
Simulated transient response for the tri-state TTL inverter.

the input was set to 10 times the frequency of the enable. The results of the transient simulation are provided in Figure 13.26. With logic one applied at the enable, normal inverter operation is obtained; with logic zero applied at the enable, the high Z state is obtained at the output. As with the previous TTL example, a resistor divider network was added to the output circuitry to show this clearly. Therefore, with the totem pole output floating, the voltage at the output node goes to $V_{CC}/2$.

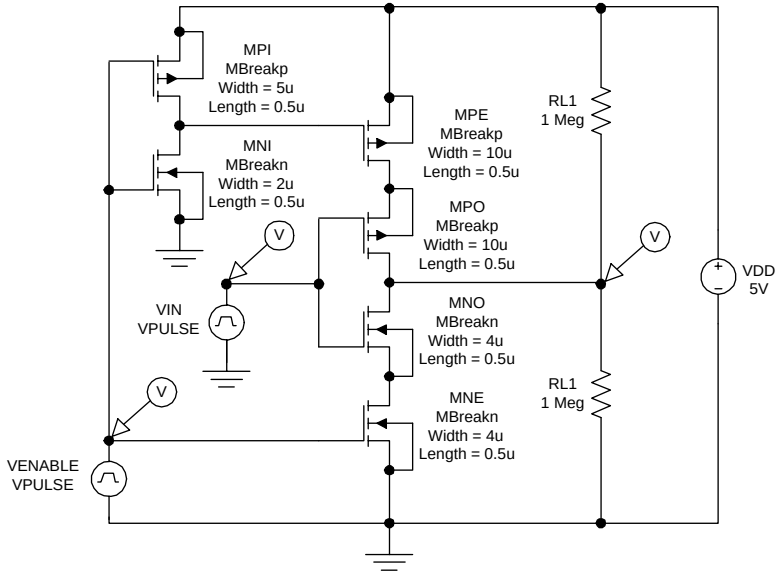


FIGURE 13.25
Tri-state CMOS inverter circuit.

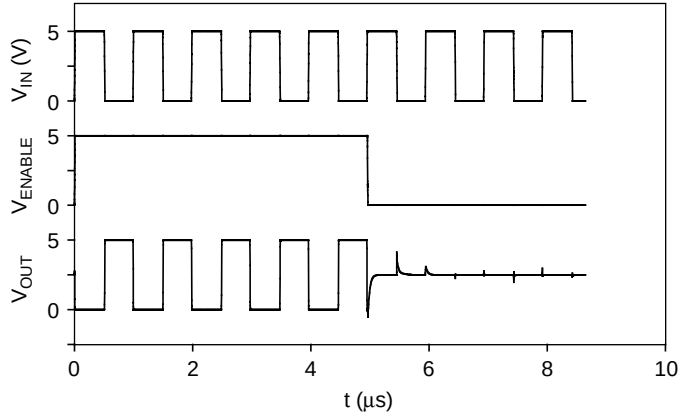


FIGURE 13.26
Simulated transient response for the tri-state CMOS inverter.

13. 7 Summary

The design of digital systems involves the consideration of issues outside of the individual gate circuits. These include interfacing, wired logic, and tri-

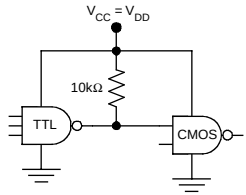
state logic. Most systems include digital gates from different logic families; this necessitates interfacing. The interfacing between ECL and TTL requires level translator circuits that translate between the voltage levels required by these logic families. In other cases, voltage translation is not necessary but current loading must be considered (as in CMOS-to-TTL interfacing).

Often, the outputs of several gates must be connected to a common node, which can result in wired logic. If the individual circuits have active pull-down circuitry, wired logic results in the ANDing of the individual signals. If the individual circuits have active pull-up circuitry, wired logic results in the OR function. Wired logic may not be used with circuits that have active pull-up and active pull-down.

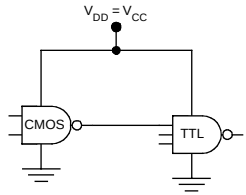
In bus-organized systems, it is often necessary to connect the outputs of many gates to a single node even though the circuits have active pull-up and active pull-down. In such cases, tri-state gates must be used to avoid conflicts. In this way, a single gate may drive the common node while the other gates are "tri-stated" or in the "high Z" state. Tri-state gates may be constructed by using transmission gates or by internal modifications to two-state gate circuits.

INTERFACING BETWEEN DIGITAL LOGIC CIRCUITS QUICK REFERENCE

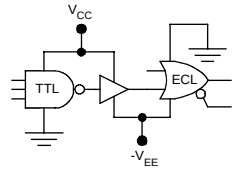
Interfacing



TTL to CMOS with pull-up resistor



CMOS to TTL

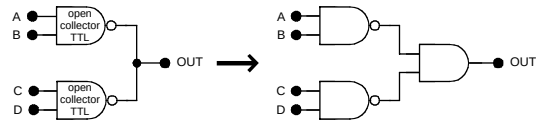


TTL to ECL with level translator

Most digital systems combine gates from different circuit families; this necessitates interfacing. Often, the outputs of several gates are connected to a common node. This can result in wired logic. In other cases, transmission gates or tri-state gates must be used to avoid conflicts on common nodes such as those on data or address busses.

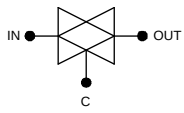
Wired Logic

Active pull-down	AND
Active pull-up	OR
Active pull-down and active pull-up	can not be used

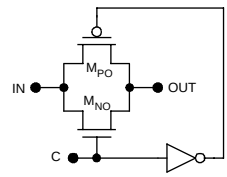


wired logic with open-collector TTL

Transmission Gates

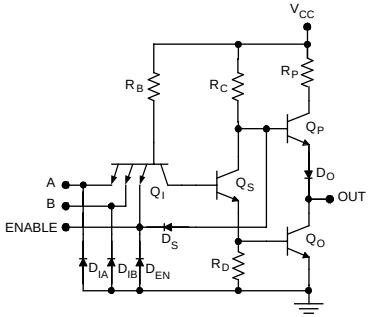


C	IN	OUT
0	X	High Z
1	0	0
1	1	1

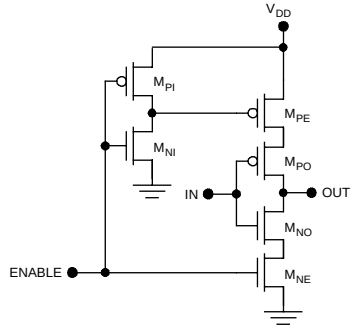


CMOS transmission gate

Tri-State Logic



Tri-state TTL NAND2 gate



Tri-state CMOS inverter

$f = 10^{-15}$ $p = 10^{-12}$ $n = 10^{-9}$ $\mu = 10^{-6}$ $m = 10^{-3}$ $k = 10^{3}$ $M = 10^{6}$ $G = 10^{9}$

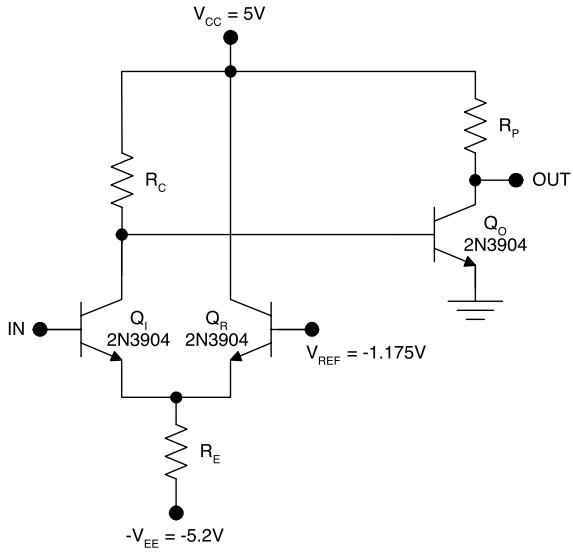


FIGURE 13.27
ECL-to-TTL level translator circuit (L13.1).

Laboratory Exercises

L13.1. For the ECL-to-TTL level translator circuit illustrated in Figure 13.27, choose the resistor values so that the average DC dissipation of the circuit is 50 mW.

1. Using SPICE, determine and plot the voltage transfer characteristic for the level translator.
2. Using SPICE, determine the average DC power dissipation for the circuit.
3. Using SPICE, determine the propagation delays for the level translator.
4. Build the circuit.
5. Using the x–y function of an oscilloscope or virtual instrument, measure and plot the voltage transfer characteristic.
6. Experimentally determine the average DC dissipation.

L13.2. For the ECL-to-TTL level translator circuit depicted in Figure 13.28, choose the resistor values so that the average DC dissipation of the circuit is 75 mW.

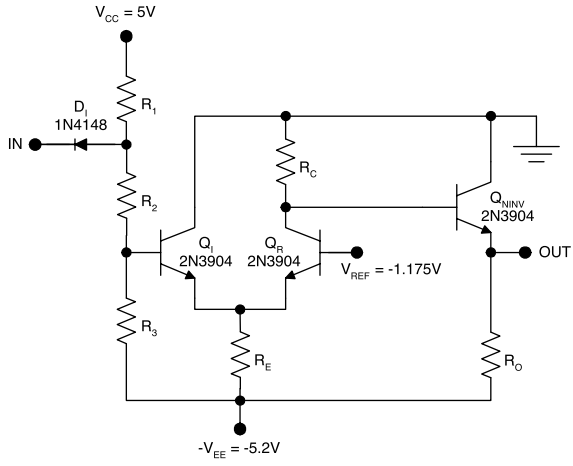


FIGURE 13.28
ECL-to-TTL level translator circuit (L13.2).

1. Using SPICE, determine and plot the voltage transfer characteristic for the level translator.
2. Using SPICE, determine the average DC power dissipation for the circuit.
3. Using SPICE, determine the propagation delays for the level translator.
4. Build the circuit.
5. Using the x–y function of an oscilloscope or virtual instrument, measure and plot the voltage transfer characteristic.
6. Experimentally determine the average DC dissipation.

L13.3. Using commercially available semiconductor devices, design a circuit to interface between 3-V CMOS and 6-V CMOS. The requirements for the circuit are: $V_{OL} < 0.3$ V, $V_{OH} > 5.7$ V, and $P < 10$ mW.

1. Using SPICE, verify that the circuit meets all specifications.
2. Build the circuit and experimentally verify that it meets all specifications.

L13.4. Use commercially available enhancement-type MOSFETs to build a CMOS transmission gate based on Figure 13.29. Obtain the data sheets for the devices from the manufacturer's Web pages. The on-resistance of the transmission gate may be defined as

$$R_{ON} = \left| \frac{dV_{\text{OUT}}}{dI_{\text{OUT}}} \right|_{C=V_{DD}}.$$

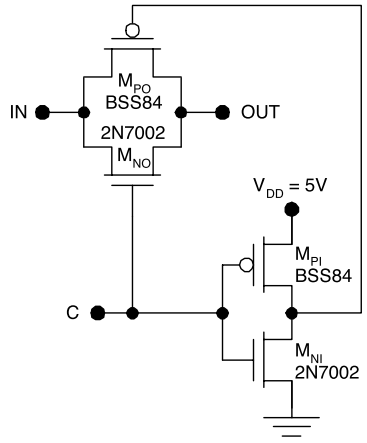


FIGURE 13.29
CMOS transmission gate (L13.4).

1. Using hand calculations, determine and plot the on-resistance of the transmission gate as a function of V_{IN}.
2. Using SPICE, determine and plot the on-resistance of the transmission gate as a function of V_{IN}.
3. Build the transmission gate and make measurements to determine the on-resistance as a function of V_{IN}.

L13.5. Design a 5-V CMOS to 3.3-V CMOS level shifter using commercially available enhancement-type MOSFETs. Use SPICE to determine the voltage transfer characteristic for the circuit and thus verify its level-shifting performance. Build the circuit and experimentally determine the voltage transfer characteristic using the x–y function of an oscilloscope or virtual instrument.

L13.6. Design a 2.5-V CMOS to 3.3-V CMOS level shifter using commercially available enhancement-type MOSFETs. Use SPICE to determine the voltage transfer characteristic for the circuit and thus verify its level-shifting performance. Build the circuit and experimentally determine the voltage transfer characteristic using the x–y function of an oscilloscope or virtual instrument.

L13.7. Use commercially available enhancement-type MOSFETs to realize the tri-state CMOS NAND2 gate as shown in Figure 13.30. Obtain data sheets for the devices from the manufacturer's Web pages.

1. Design a resistive load circuit that will allow clear observation of the high Z state, but will not unduly affect the normal operation of the gate. Using SPICE, apply square wave inputs with the frequencies related as follows: $f_A = 2f_B = 4f_{ENABLE}$ and verify the function of the gate and the load circuit.

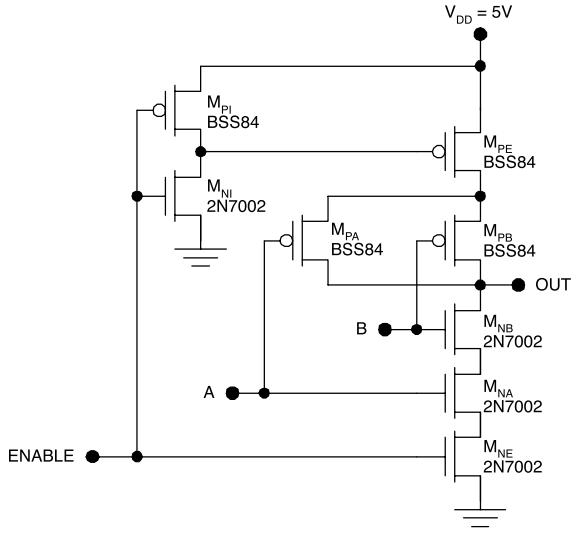


FIGURE 13.30
Tri-state CMOS NAND2 gate (L13.7).

2. Build the circuit using the load designed in (a). Experimentally verify the operation of the gate using square wave inputs with $f_A = 2f_B = 4f_{ENABLE}$.
3. Experimentally determine V_{IL} and V_{IH} for the ENABLE input. Do these critical voltages depend on the voltages at A and B?

Problems

P13.1. Consider the ECL-to-TTL level translator illustrated in Figure 13.31; $\beta_F = 50$ for all transistors.
1. Choose the resistor values so that the average DC dissipation is 30 mW.
2. Calculate and plot the voltage transfer characteristic.

P13.2. Consider the ECL-to-TTL level translator depicted in Figure 13.32; $\beta_F = 50$ for all transistors.
1. Choose the resistor values so that the average DC dissipation is 45 mW.
2. Calculate and plot the voltage transfer characteristic.

P13.3. Design a circuit to interface from 5-V CMOS to 1.5-V CMOS. Assume 0.25-μm technology (all gate lengths are 0.25 μm) and that $V_T = 0.3$ V for all circuits. Determine and plot the voltage transfer characteristic for the interface circuit.

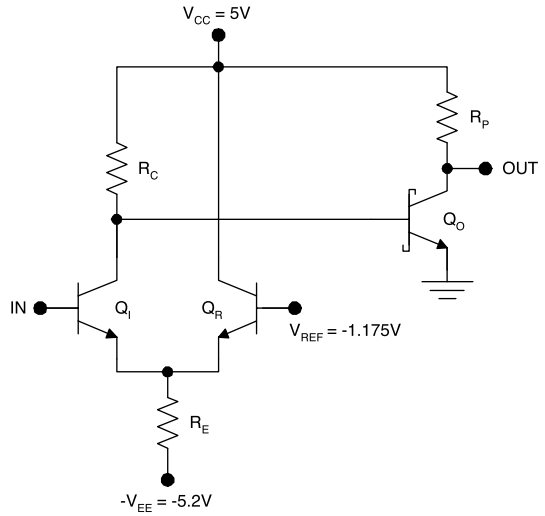


FIGURE 13.31
ECL-to-TTL level translator (P13.1).

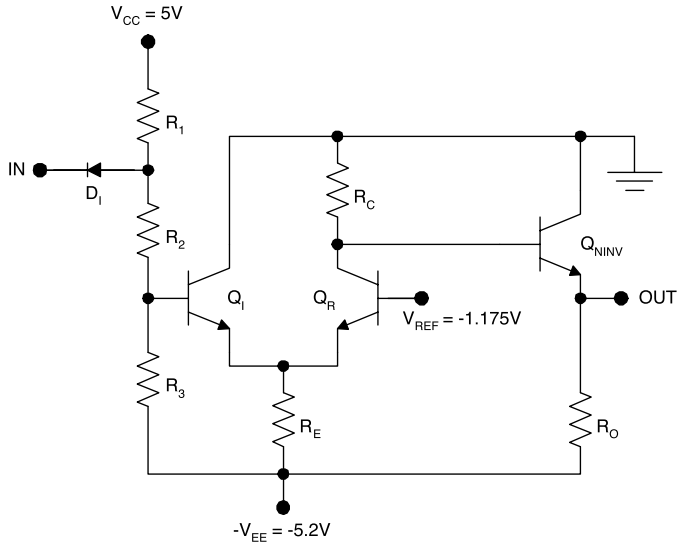


FIGURE 13.32
ECL-to-TTL level translator (P13.2).

P13.4. Design a circuit to interface from 1.5-V CMOS to 3.3-V CMOS. Assume 0.25-μm technology (all gate lengths are 0.25 μm) and that V_T = 0.3 V for all circuits. Determine and plot the voltage transfer characteristic for the interface circuit.

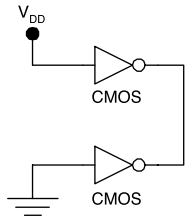


FIGURE 13.33
Two CMOS inverters connected in an attempt at wired logic (P13.5).

P13.5. Suppose that two CMOS inverters are connected in an inappropriate attempt at wired logic, as shown in Figure 13.33. K = 0.5 mA/V^2, V_T = 0.5 V, and V_{DD} = 3.3 V.

1. Determine the resulting value of supply current if the input to one gate is grounded while the input of the other gate is connected to V_{DD}.
2. Compare this level of current to the peak current under normal operation of the CMOS gate.

References

1. F100K ECL dual rail translators, Fairchild Semiconductor application note AN-784, www.fairchildsemi.com, 2000.
2. Wang, W.-T., Ker, M.-D., Chiang, M.-C., Chen, C.-H., Level shifters for high-speed 1-V to 3.3-V interfaces in a 0.13-μm Cu-interconnection/low-k CMOS technology, *Proc. 2001 Int. Symp. VLSI Technol., Syst. Appl.*, 307, 2001.
3. Steyaert, M.S.J., Bijker, W., Vorenkamp, P., and Sevenhans, J., ECL–CMOS and CMOS–ECL interface in 1.2-μm CMOS for 150-MHz digital ECL data transmission systems, *IEEE J. Solid-State Circuits*, 26, 18, 1991.
4. Steyaert, M., Bijker, W., Vorenkamp, P., and Sevenhans, J., A full 1.2-μm CMOS ECL–CMOS–ECL converter with subnanosecond settling times, *Proc. 1990 IEEE Custom Integrated Circuits Conf.*, 1, 1990.
5. Boudon, G., Wallart, F., and Maillart, E., Internal ECL–BiCMOS translator circuits in half micron technology, Proc. *1989 IEEE Int. Conf. Computer Design: VLSI Computers Processors*, 314, 1989.
6. Pedersen, M. and Metz, P., A CMOS to 100-K ECL interface circuit, *Dig. Tech. Papers 36th IEEE Int. Solid-State Circuits Conf.*, 226, 1989.
7. Gabara, T.J. and Thompson, D.W., A 200-MHz 100-K ECL output buffer for CMOS ASICs, *Proc. 3rd IEEE ASIC Semin. Exhibit*, P8/5.1, 1990.
8. Bass, A. and Eyck, T.T., Fast ECL-to-CMOS and CMOS-to-ECL translators for an 0.8-μm BiCMOS gate array, *Proc. 3rd IEEE ASIC Semin. Exhibit*, P15/5.1, 1990.
9. Sevenhans, J., Bijker, W., Vorenkamp, P., and Steyaert, M., Performance of ECL-compatible 75-Ω line driver/receiver realized in a 1.2-μm CMOS technology, *Electron. Lett.*, 26, 764, 1990.

10. Schou, G., Cherel, J., Perea, E.H., Danckaert, J.-Y., Dean, T., and Chaplard, J., Fully ECL-compatible GaAs standard-cell library, *IEEE J. Solid-State Circuits*, 23, 676, 1988.
11. Mansoorian, B., Ozguz, V., and Esener, S., Diode-biased AC-coupled ECL-to-CMOS interface circuit, *IEEE J. Solid-State Circuits*, 28, 397, 1993.
12. Gu, R.X. and Elmasry, M.I., High-speed dynamic reference voltage (DRV) CMOS/ECL interface circuits, *IEEE J. Solid-State Circuits*, 29, 1282, 1994.
13. FACT design considerations, Fairchild Semiconductor document MS-539, www.fairchildsemi.com, 1999.
14. Nakase, Y., Suzuki, H., Makino, H., Shinohara, H., and Mashiko, K., A BiCMOS wired-OR logic, *IEEE J. Solid-State Circuits*, 30, 622, 1995.
15. CMOS quad bilateral switch, Texas Instruments CD4066B data sheet, www.ti.com, 2002.
16. www.intel.com (Intel).
17. www.motorola.com (Motorola).
18. www.amd.com (Advanced Micro Devices).
19. www.ibm.com (IBM).
20. www.lucent.com (Lucent Technologies).
21. www.fairchildsemi.com (Fairchild Semiconductor).
22. www.national.com (National Semiconductor).
23. www.cadence.com (Cadence).

14

Interconnect

14.1 Introduction

Interconnect refers to the metal wires that make electrical connections between the transistors on a die. Aluminum and copper are commonly used as interconnect metals, but doped polysilicon is also used in some situations. Interconnect has attracted increasing attention over the last few years because of its growing influence on the overall performance of digital integrated circuits.[1–4] This has come about by the scaling of device dimensions coupled with the trend to larger die sizes. Currently, microprocessors contain about 1 km of interconnect for every square centimeter of die area.

Interconnect introduces parasitic capacitances, resistances, and inductances that can degrade the overall performance significantly. Interconnect capacitances present considerable loading to CMOS circuits, increasing the propagation delays and the switching dissipation. The trend toward larger die sizes has also necessitated using longer interconnects on the chip. In such long interconnects, the parasitic resistance must be considered as well as the capacitance. The associated RC delays further degrade the overall circuit performance; in some special cases, the interconnects can be so long that they must be treated as lossy RLC transmission lines rather than RC networks. Then care must be taken to terminate the transmission lines to avoid reflections.

In all of the situations outlined here, the parasitics tend to degrade performance. In itself, the additional capacitive loading is responsible for increased propagation delays and power. In some digital integrated circuits, interconnects can account for one quarter of the total dissipation. In addition, the parasitic capacitances between interconnects tend to introduce interference, called cross talk. As a consequence, the propagation delay for an interconnect becomes a function of the signals on neighboring interconnects. This situation is highly undesirable because it makes performance predictions difficult.

This chapter will discuss the parasitic resistances, capacitances, and inductances of interconnect. The lumped, distributed, and transmission line models

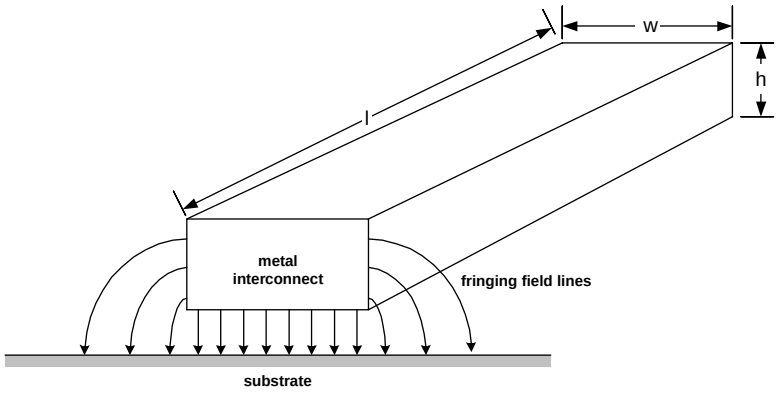


FIGURE 14.1
Estimation of the interconnect capacitance.

for interconnect will be outlined and rules developed for determination of the most appropriate model. Special problems in interconnect design will be discussed from the point of view of minimizing cross talk and optimizing performance. Finally, SPICE models for interconnect will be described with some examples.

14.2 Capacitance of Interconnect

Capacitance is the most important parasitic introduced by interconnect wires. Some of this capacitance appears between the interconnect and ground; another component, called *interwire capacitance*, appears between wires on a single level. The *interlevel capacitance* involves capacitances between the interconnect and wires on other levels. In an integrated circuit utilizing eight to ten levels of interconnect, accurate modeling of the capacitive effects is rather complex and beyond the scope of this book.[5–7] However, it is useful to consider some simple models to understand the general behavior of interconnect capacitances.

Consider a single level of interconnect with a rectangular cross section* routed over a semiconductor substrate with an intermediate dielectric layer as shown in Figure 14.1. The two components of the capacitance between this *microstripline* and the substrate (ground plane) are the parallel plate capacitance associated with the parallel field lines directly underneath the interconnect and the fringing field capacitance.

Including both components, the effective capacitance between the interconnect and the substrate is approximately

* The rectangular cross sections for wires are dictated by the deposition, lithography, and etching processes.

TABLE 14.1

Typical Values of Relative Permittivity for Dielectric Materials

Dielectric	Relative Permittivity
SiO_2	3.9
Fluorosilicate glass	3.6
Carbon-doped oxide	2.7–2.9
Si-based polymers	2.2–2.6

$$C \approx \underbrace{\frac{\varepsilon_{DI} l w}{t_{DI}}}_{\substack{\text{parallel plate} \\ \text{capacitance}}} + \underbrace{\frac{\varepsilon_{DI} l}{\log\left(1 + t_{DI}/h\right)}}_{\substack{\text{fringing field} \\ \text{capacitance}}}, \tag{14.1}$$

where

ε_{DI} = permittivity of the dielectric
t_{DI} = thickness of the dielectric between the interconnect and the substrate
w = width of the interconnect
h = height of the interconnect
l = length of the interconnect

The capacitance per unit length of interconnect is given by

$$c = \frac{\varepsilon_{DI} w}{t_{DI}} + \frac{\varepsilon_{DI}}{\log\left(1 + t_{DI}/h\right)}. \tag{14.2}$$

Typical relative permittivities for dielectric materials are shown in Table 14.1. Chemical vapor-deposited (CVD) silicon dioxide films are commonly used as interlayer dielectrics. Recently, CVD fluorosilicate glass has been applied for high-performance integrated circuits. Fluorosilicate glass reduces the permittivity, and therefore the parasitic capacitances, by about 10%. However, this is not adequate for the next few generations of integrated circuits. Instead, other *low-k* dielectrics* such as CVD carbon-doped oxide or spin-on polymers must be used.[8–14]

Example 14.1

Estimate the capacitance to ground per unit length for a polysilicon interconnect, 0.25 μm wide and 0.25 μm thick, on a 0.5-μm thick layer of silicon dioxide ($\varepsilon_r = 3.9$).

* Sometimes the relative permittivity is denoted by the Greek letter κ. Therefore, materials with low permittivity are often called "low-k dielectrics."

Solution. The capacitance to the substrate per unit length is

$$c = \frac{\varepsilon_{DI} w}{t_{DI}} + \frac{\varepsilon_{DI}}{\log(1 + t_{DI}/h)}$$

$$= \frac{(3.9)(8.85 \times 10^{-14}\ \text{F/cm})(0.25 \times 10^{-4}\ \text{cm})}{0.5 \times 10^{-4}\ \text{cm}} + \frac{(3.9)(8.85 \times 10^{-14}\ \text{F/cm})}{\log(1 + 0.5/0.25)}$$

$$= 1.72 \times 10^{-13}\ \text{F/cm} + 7.2 \times 10^{-13}\ \text{F/cm} = 0.89\ \text{pF/cm}$$

These numbers are representative of a polysilicon wire running over the field oxide used in a 0.25-μm CMOS process. Because the fringing field capacitance is dominant, the capacitance per unit length is a weak function of the polysilicon width.

Example 14.2

Estimate the capacitance to ground per unit length for an aluminum interconnect, 0.375 μm wide and 0.25 μm thick, on a 0.5-μm thick layer of silicon dioxide ($\varepsilon_r = 3.9$).

Solution. The capacitance to ground per unit length is

$$c = \frac{\varepsilon_{DI} w}{t_{DI}} + \frac{\varepsilon_{DI}}{\log(1 + t_{DI}/h)}$$

$$= \frac{(3.9)(8.85 \times 10^{-14}\ \text{F/cm})(0.375 \times 10^{-4}\ \text{cm})}{0.5 \times 10^{-4}\ \text{cm}} + \frac{(3.9)(8.85 \times 10^{-14}\ \text{F/cm})}{\log(1 + 0.5/0.25)}$$

$$= 2.6 \times 10^{-13}\ \text{F/cm} + 7.2 \times 10^{-13}\ \text{F/cm} = 0.98\ \text{pF/cm}$$

These numbers are representative of a level-one metal interconnect used in a 0.25-μm CMOS process. As for the case of polysilicon, the fringing field capacitance is dominant because of the nearly square cross section for the interconnect. It is possible to neglect the fringing field capacitance only for very wide interconnects with $w \gg h$.

Realistic estimates of the total capacitance per unit length must include the interwire and interlevel contributions. As a rough rule of thumb, the total capacitance per unit length for the metal-one layer is two times the capacitance to ground. The total capacitance per unit length is relatively constant from one level to the next for a given width of interconnect. Therefore, the increase in capacitance per unit length for the upper levels of metal is dictated by the increase in metal width. More detailed design rules are available for the different CMOS processes but no attempt will be made to catalog them here.

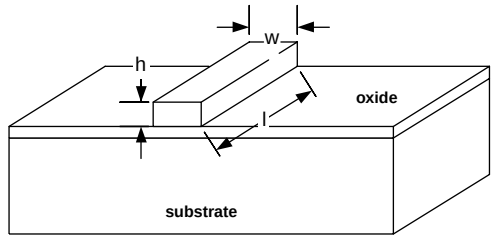


FIGURE 14.2
Interconnect with a rectangular cross section for the estimation of resistance.

TABLE 14.2

Typical Resistivities for Interconnect Materials

Interconnect Material	Resistivity (Ωcm)
P-type polysilicon	0.02
N-type polysilicon	0.01
Aluminum	2.8×10^{-6}
Copper	1.7×10^{-6}

14.3 Resistance of Interconnect

Consider a metal interconnect with a rectangular cross section as shown in Figure 14.2. The low-frequency resistance of this interconnect may be calculated as

$$R = \frac{l}{w}\frac{\rho}{h}, \tag{14.3}$$

where ρ is the resistivity of the interconnect material and l, w, and h are the length, width, and thickness of the interconnect, respectively. The resistance per unit length is

$$r = \frac{\rho}{wh}. \tag{14.4}$$

Typical resistivities for interconnect materials are shown in Table 14.2.

Example 14.3

Estimate the low-frequency resistance per unit length for an aluminum interconnect, 0.5 μm thick and 0.5 μm wide.

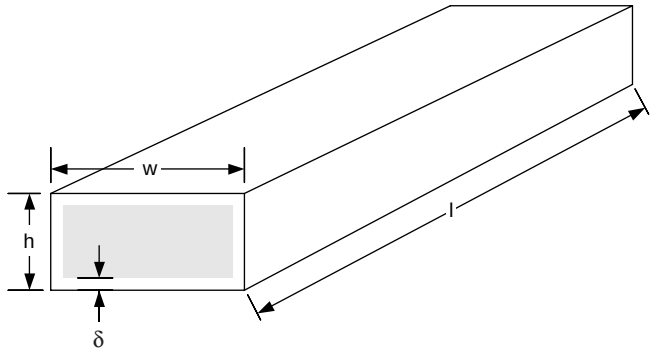


FIGURE 14.3
The skin effect in interconnect.

Solution. Assuming the interconnect takes on the bulk resistivity of aluminum,

$$r = \frac{\rho}{wh} = \frac{\left(2.8\times10^{-6}\,\Omega\text{cm}\right)}{\left(0.5\times10^{-4}\,\text{cm}\right)\left(0.5\times10^{-4}\,\text{cm}\right)} = 1.12\ \text{k}\Omega/\text{cm}\,.$$

The magnitude of this value suggests that the interconnect resistance may be neglected only for short wires. However, the choice of including or neglecting the interconnect resistance is influenced by the interconnect capacitance, as will be shown later.

At high frequencies, the resistance becomes frequency dependent due to the *skin effect*.[15] This is because, at high frequencies, the current flow becomes concentrated near the outer surfaces of the interconnect. This effect can be quantized by the *skin depth*, which is the depth at which the current density falls to 1/e times the surface value. This skin depth is given by

$$\delta = \sqrt{\frac{\rho}{\pi f \mu}}\,, \tag{14.5}$$

where f is the frequency and μ is the permeability of the interconnect.

The skin effect increases the effective resistance of the interconnect. This effect can be accounted for approximately by assuming that the cross sectional area is reduced to that within the skin depth, as shown in Figure 14.3. Based on this figure, the effective resistance of the interconnect is approximately

$$R = \frac{l\rho}{\delta\left[2w + 2(h - 2\delta)\right]}\,, \tag{14.6}$$

and the resistance per unit length is

$$r = \frac{\rho}{\delta\left[2w + 2(h - 2\delta)\right]}, \tag{14.7}$$

Example 14.4

Estimate the skin depth in copper interconnect at a frequency of 100 GHz.

Solution. It is assumed that the copper and the dielectric in which it is embedded have permeabilities equal to that of a vacuum. At 100 GHz, the skin depth is

$$\delta(Cu) = \sqrt{\frac{\rho}{\pi f \mu}} = \sqrt{\frac{\left(1.7 \times 10^{-6}\ \Omega\text{cm}\right)}{\pi\left(10^{11}\ \text{Hz}\right)\left(4\pi \times 10^{-9}\ \text{H/cm}\right)}} = 0.21 \times 10^{-4}\ \text{cm} = 0.21\ \mu\text{m}\ .$$

Therefore, the skin effect will be important at 100 GHz in Cu interconnects with widths greater than twice this value.

Example 14.5

Estimate the skin depth vs. frequency in copper interconnect.

Solution. It is assumed that the copper and the dielectric in which it is embedded have permeabilities equal to that of a vacuum. The skin depth is

$$\delta(Cu) = \sqrt{\frac{\rho}{\pi f \mu}} = \sqrt{\frac{\left(1.7 \times 10^{-6}\ \Omega\text{cm}\right)}{\pi f\left(4\pi \times 10^{-9}\ \text{H/cm}\right)}} = 0.21\ \mu\text{m}\sqrt{\frac{100\ \text{GHz}}{f}}\ .$$

The results are shown in Figure 14.4.

Example 14.6

Consider interconnects 1 μm thick and 1 μm wide. Compare the resistances per unit length for aluminum and copper interconnect, assuming a frequency of 100 GHz.

Solution. It is assumed that these metals and the dielectric in which they are embedded have permeabilities equal to that of a vacuum. At 100 GHz, the skin depths are

$$\delta(Cu) = \sqrt{\frac{\rho}{\pi f \mu}} = \sqrt{\frac{\left(1.7 \times 10^{-6}\ \Omega\text{cm}\right)}{\pi\left(10^{11}\ \text{Hz}\right)\left(4\pi \times 10^{-9}\ \text{H/cm}\right)}} = 0.21 \times 10^{-4}\ \text{cm} = 0.21\ \mu\text{m}$$

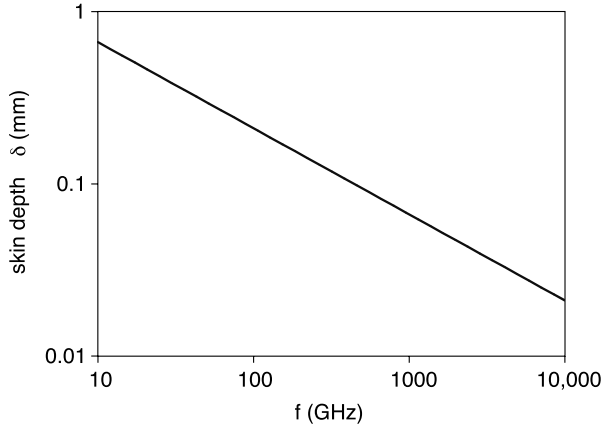


FIGURE 14.4
Skin depth vs. frequency for Cu interconnect.

and

$$\delta(Al) = \sqrt{\frac{\rho}{\pi f \mu}} = \sqrt{\frac{\left(2.8\times10^{-6}\ \Omega\text{cm}\right)}{\pi\left(10^{11}\ \text{Hz}\right)\left(4\pi\times10^{-9}\ \text{H/cm}\right)}} = 0.27\times10^{-4}\ \text{cm} = 0.27\ \mu\text{m}\,.$$

On a per-unit length basis, the approximate interconnect resistances are

$$r(Cu) = \frac{\left(1.7\times10^{-6}\ \Omega\text{cm}\right)}{\left(0.21\times10^{-4}\ \text{cm}\right)\left[2\left(1.0\times10^{-4}\ \text{cm}\right)+2\left(1.0\times10^{-4}\ \text{cm}-0.42\times10^{-4}\ \text{cm}\right)\right]}$$

$$= 260\ \ \Omega/\text{cm}$$

and

$$R(Al) = \frac{\left(2.7\times10^{-6}\ \Omega\text{cm}\right)}{\left(0.27\times10^{-4}\ \text{cm}\right)\left[2\left(1.0\times10^{-4}\ \text{cm}\right)+2\left(1.0\times10^{-4}\ \text{cm}-0.52\times10^{-4}\ \text{cm}\right)\right]}\,.$$

$$= 340\ \Omega$$

Although the bulk resistivity of Al is 60% higher than that of Cu, the interconnect resistance is only 30% higher in this example. Therefore, in these wide interconnects, the skin effect negates some of the advantage of using the better conductor (Cu).

The preceding analysis is approximate because it assumes a sinusoidal current waveform and the abrupt confinement of the current within the skin depth; nevertheless, three important conclusions can be drawn from this

approximate analysis. First, the skin effect is important in high-performance digital integrated circuits utilizing clock frequencies in the GHz range. Second, the skin effect is more important in wide interconnects (much wider than the skin depth). Third, the skin effect negates some of the advantage associated with using a lower-resistivity conductor because the skin depth is less in a lower-resistance conductor.

14.4 Inductance of Interconnect

The inductance of an interconnect may be estimated most easily, if the capacitance is known, by using the approximate expression

$$L \approx \frac{\varepsilon_{DI}\mu_{DI}}{C}, \tag{14.8}$$

where ε_{DI} and μ_{DI} are the permittivity and permeability of the interconnect, respectively. If the interconnect is sandwiched within layers of different dielectrics, average values of the permeability and permittivity may be used in an approximate analysis. Similarly, the inductance per unit length may be found from

$$l \approx \frac{\varepsilon_{DI}\mu_{DI}}{c}. \tag{14.9}$$

Example 14.7

Estimate the inductance per unit length for a copper interconnect, 0.375 μm wide and 0.25 μm thick on a 1.0-μm thick layer of silicon dioxide ($\varepsilon_r = 3.9$).

Solution. From Example 14.2, the capacitance to ground per unit length of this interconnect is 0.98 pF/cm. Therefore, the inductance per unit length is approximately

$$l = \frac{\varepsilon_{DI}\mu_{DI}}{c} = \frac{(3.9)\left(8.85\times10^{-14}\ \text{F/cm}\right)\left(4\pi\times10^{-9}\ \text{H/cm}\right)}{0.98\times10^{-12}\ \text{F/cm}} = 4.4\times10^{-9}\ \text{H/cm}.$$

This inductance is very small and can be neglected under most circumstances. Thus, although parasitic inductances associated with package pins are important, the parasitic inductances associated with interconnect are important only for long wires and very high-frequency operation. Increasingly, however, designers consider the inductance or even invoke transmission line models for interconnect.[16–20]

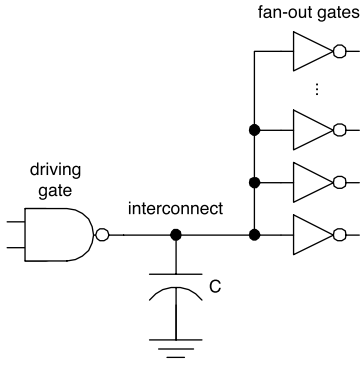


FIGURE 14.5
Use of a lumped capacitance to model an interconnect.

14.5 Lumped Capacitance Model

It is reasonable to model short runs of interconnect using a *lumped capacitance model* as shown in Figure 14.5. This approach is valid even for a branching interconnect that feeds a number of fan-out gates, as long as the resistance can be neglected and the capacitance is calculated based on the total length of interconnect. The primary effect of the interconnect is to increase the effective load capacitance, which, in turn, increases the propagation delay and the dissipation for the driving gate.

14.6 Distributed Models

In longer runs of interconnect, the parasitic resistances become important. Therefore, distributed RC models should be used in these situations. Consider a general branching interconnect modeled by an N*th* order distributed network as shown in Figure 14.6. This network exhibits N time constants and the exact analysis is rather complex. However, it is usually sufficient to consider only the dominant time constant (the *Elmore delay*), which greatly simplifies the analysis.[20–23] Based on this approach, the propagation delay from the driving gate (at node zero) to the i*th* node is approximately

$$t_{Pi} \approx \ln(2) \sum_{k=1}^{N} C_k R_{ik} \ , \tag{14.10}$$

where R_{ik} is the shared path resistance for nodes i and k (i.e., the resistance common to the path from the driving gate to node i and the path from the driving gate to node k), given by

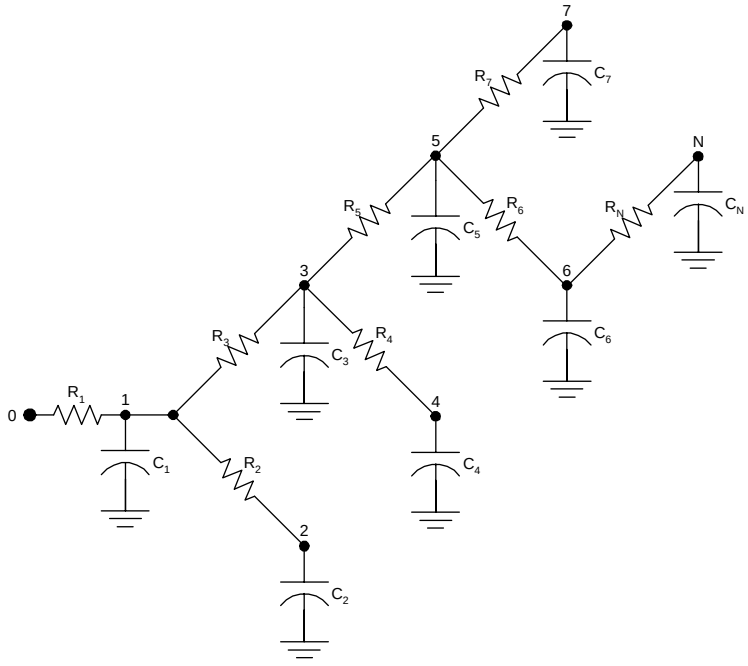


FIGURE 14.6
Distributed RC model for a branching interconnect.

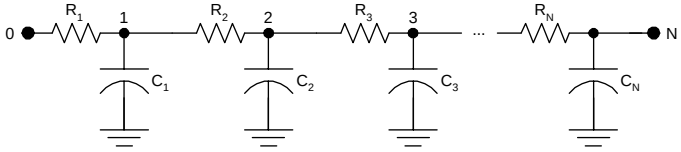


FIGURE 14.7
Distributed RC model for an interconnect.

$$R_{ik} = \sum R_j \in \left[path(0 \to i) \cap \left(path(0 \to k) \right) \right]. \tag{14.11}$$

A special case is the straight (nonbranching) interconnect shown in Figure 14.7. Here, the Elmore approach yields an approximate end-to-end propagation delay of

$$\begin{aligned} t_p &\approx \ln(2)\left[C_1 R_1 + C_2 \left(R_1 + R_2 \right) + C_3 \left(R_1 + R_2 + R_3 \right) + ... \right] \\ &= \ln(2) \sum_{i=1}^{N} C_i \sum_{j=1}^{i} R_j \end{aligned} \tag{14.12}$$

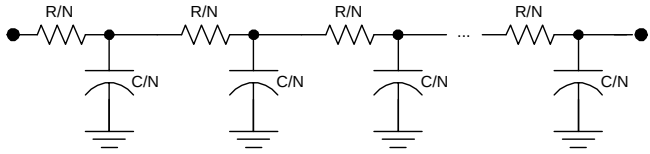


FIGURE 14.8
Evenly distributed RC model for an interconnect.

A situation of special interest is the nonbranching interconnect with evenly distributed resistance and capacitance, as shown in Figure 14.8.

The propagation delay for the evenly distributed RC network, divided into N segments, is

$$t_P \approx \ln(2)\left[\frac{RC}{N^2}+\frac{2RC}{N^2}+\frac{3RC}{N^2}+\ldots+\frac{NRC}{N^2}\right]=\ln(2)RC\frac{N+1}{2N}. \quad (14.13)$$

If the number of segments is increased arbitrarily, the propagation delay asymptotically reaches the limiting value

$$t_P \approx \ln(2)\lim_{N\to\infty}\left[RC\frac{N+1}{2N}\right]=\ln(2)\frac{RC}{2}=\ln(2)\frac{rcl^2}{2}, \quad (14.14)$$

where r is the resistance per unit length, c is the capacitance per unit length, and l is the length of the interconnect. Therefore, *the delay associated with an evenly distributed RC interconnect is proportional to the square of its length*. As a consequence, long signal lines are usually broken up into a series of shorter runs separated by buffer gates called *repeaters*.[24]

The choice of a lumped capacitance model or the distributed rc model for the interconnect may be made after comparing the propagation delay for the driving gate and the time constant for the interconnect:

$$t_{P,driver} > \ln(2)\frac{rcl^2}{2} \quad \Rightarrow \quad \text{lumped C model}$$

$$t_{P,driver} < \ln(2)\frac{rcl^2}{2} \quad \Rightarrow \quad \text{distributed RC model}$$

In terms of the interconnect length, this rule of thumb may be restated as:

$$l < \sqrt{\frac{t_{P,driver}}{rc\ln(2)/2}} \quad \Rightarrow \quad \text{lumped C model}$$

$$l > \sqrt{\frac{t_{P,driver}}{rc\ln(2)/2}} \quad \Rightarrow \quad \text{distributed RC model}$$

14.7 Transmission Line Model

Very long wires must be treated as transmission lines. At the present time, this sometimes applies to metal runs on printed circuit boards but only rarely applies to interconnects on the integrated circuit. This situation may change as circuit speeds are increased while die sizes continue to increase, however.

In a transmission line, the signal travels as a wave at the speed of light, given by

$$v = \frac{c_0}{\sqrt{\varepsilon_r \mu_r}}, \tag{14.15}$$

where

c_0 = speed of light in a vacuum = 3.00×10^{10} cm/s
ε_r = relative permittivity of the transmission line medium
μ_r = relative permeability of the transmission line medium

Example 14.8

Estimate the time of flight for a signal crossing a 1-cm integrated circuit by a copper transmission line embedded in silicon dioxide.

Solution. For silicon dioxide, the relative permittivity is 3.9 and the relative permeability is ~1.0. Thus, the speed of propagation is

$$v = \frac{c_0}{\sqrt{\varepsilon_r \mu_r}} = \frac{3.0 \times 10^{10}\ \text{cm/s}}{\sqrt{(3.9)(1.0)}} = 1.52 \times 10^{10}\ \text{cm/s}$$

and the time of flight is

$$t_{flight} = \frac{1\ \text{cm}}{1.52 \times 10^{10}\ \text{cm/s}} = 6.6 \times 10^{-11}\ \text{s} = 66\ \text{ps}.$$

This time of flight will be comparable to the clock period in a 15-GHz processor. It is clear from this example that clock distribution presents a special problem in large digital integrated circuits.

Consider a general two-wire transmission line as shown in Figure 14.9. The equations governing the general two-wire transmission line are

$$\frac{\partial^2 V}{\partial x^2} = -(r + j2\pi fl)(g + j2\pi fc)V \tag{14.16}$$

and

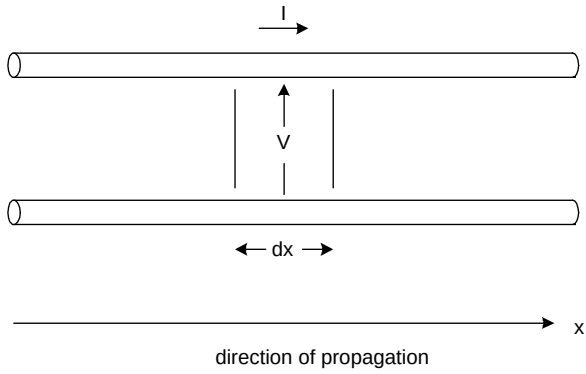


FIGURE 14.9
Two-wire transmission line.

$$\frac{\partial^2 I}{\partial x^2} = -\left(r + j2\pi fl\right)\left(g + j2\pi fc\right)I\,, \tag{14.17}$$

where

I = current at the point x
V = voltage at the point x
r = series resistance per unit length
l = series inductance per unit length
g = parallel conductance per unit length
c = parallel capacitance per unit length
f = frequency of the propagating wave

In the special case of a lossless two-wire transmission line ($r = 0$, $g = 0$), the equation governing the propagation of a wave down the transmission line is

$$\frac{\partial^2 V}{\partial x^2} = lc\frac{\partial^2 V}{\partial t^2} = \frac{1}{v^2}\frac{\partial^2 V}{\partial t^2}\,. \tag{14.18}$$

Thus, a step in voltage applied at one end of the lossless transmission line will propagate down the line at a velocity, v, and without attenuation. Also, at any point in an infinite lossless transmission line,

$$\frac{V}{I} = \sqrt{\frac{l}{c}} = Z_0 \tag{14.19}$$

and the remainder of the line appears to have a real impedance, Z_0, called the *characteristic impedance* of the transmission line.

Example 14.9

Estimate the characteristic impedance for aluminum interconnect, 0.375 μm wide and 0.25 μm thick, on a 0.5-μm thick layer of silicon dioxide.

Solution. The capacitance per unit length is

$$c = \frac{\varepsilon_{DI} w}{t_{DI}} + \frac{\varepsilon_{DI}}{\log(1 + t_{DI}/h)}$$

$$= \frac{(3.9)(8.85\times10^{-14}\ \text{F/cm})(0.25\times10^{-4}\ \text{cm})}{0.5\times10^{-4}\ \text{cm}} + \frac{(3.9)(8.85\times10^{-14}\ \text{F/cm})}{\log(1+0.5/0.25)}$$

$$= 0.98\times10^{-12}\ \text{F/cm}$$

The inductance per unit length is approximately

$$l = \frac{\varepsilon_{DI}\mu_{DI}}{c} = \frac{(3.9)(8.85\times10^{-14}\ \text{F/cm})(4\pi\times10^{-9}\ \text{H/cm})}{0.98\times10^{-12}\ \text{F/cm}} = 4.4\times10^{-9}\ \text{H/cm}$$

and the characteristic impedance of the line is

$$Z_0 = \sqrt{\frac{l}{c}} = \sqrt{\frac{4.4\times10^{-9}\ \text{H/cm}}{0.98\times10^{-12}\ \text{F/cm}}} = 67\ \Omega\,.$$

Any wave propagating down a transmission line will be reflected from the end unless the transmission line is terminated by a real impedance equal to Z_0. If the line is terminated by a resistance, R_L, then the reflection coefficient (the ratio of the reflected to the incident voltage is given by

$$\rho = \frac{R_L - Z_0}{R_L + Z_0}\,. \tag{14.20}$$

The reflected wave will travel back to the driving gate, where it will be reflected once again. The presence of reflected waves will upset the integrity of signals on the line unless matched terminations are used.

As a rule of thumb, a transmission line model must be used only if the time of flight is greater than the propagation delay for the circuit. In terms of the interconnect length, l, this rule is:

$$l < \frac{t_{P,driver} c_0}{\sqrt{\varepsilon_r \mu_r}} \quad \Rightarrow \quad \text{distributed RC model}$$

$$l > \frac{t_{P,driver} c_0}{\sqrt{\varepsilon_r \mu_r}} \quad \Rightarrow \quad \text{transmission line model}$$

At the current time, it is seldom necessary to invoke the transmission line model for interconnect; however, this conclusion is likely to change with decreasing circuit propagation delays and increasing die sizes. The simple theory outlined previously applies to lossless transmission lines; however, real interconnects are lossy, causing an attenuation of the signals propagating on them.

Example 14.10

For aluminum interconnect, 0.5 μm wide and 0.5 μm thick on a 1.0-μm thick layer of silicon dioxide, determine the ranges of the length for which the lumped capacitance, distributed rc, and transmission line models are applicable. Assume the driving gate exhibits a propagation delay of 50 ps.

Solution. The resistance per unit length is

$$r = \frac{\rho}{wh} = \frac{\left(2.8\times10^{-6}\ \Omega\text{cm}\right)}{\left(0.5\times10^{-4}\ \text{cm}\right)\left(0.5\times10^{-4}\ \text{cm}\right)} = 1.12\ \ \text{k}\Omega/\text{cm}$$

and the capacitance per unit length is

$$\begin{aligned} c &= \frac{\varepsilon_{DI} w}{t_{DI}} + \frac{2\pi\varepsilon_{DI}}{\log\left(1+t_{DI}/h\right)} \\ &= \frac{(3.9)\left(8.85\times10^{-14}\ \text{F/cm}\right)\left(0.5\times10^{-4}\ \text{cm}\right)}{1.0\times10^{-4}\ \text{cm}} + \frac{(3.9)\left(8.85\times10^{-14}\ \text{F/cm}\right)}{\log(1+1.0/0.5)} \\ &= 1.72\times10^{-13}\ \text{F/cm} + 7.2\times10^{-13}\ \text{F/cm} = 0.98\ \ \text{pF/cm} \end{aligned}$$

The lumped capacitance model applies if

$$l < \sqrt{\frac{t_{P,driver}}{rc\ln(2)/2}} = \sqrt{\frac{50\times10^{-12}\ \text{s}}{\left(1120\ \ \Omega/\text{cm}\right)\left(0.98\times10^{-12}\ \text{F/cm}\right)\ln(2)/2}} = 3600\ \mu\text{m}\ ;$$

the transmission line model applies if

$$l > \frac{t_{P,driver} c_0}{\sqrt{\varepsilon_r \mu_r}} = \frac{\left(50\times10^{-12}\ \text{s}\right)\left(3.0\times10^{10}\ \text{cm/s}\right)}{\sqrt{(3.9)(1.0)}} = 7600\ \mu\text{m}\ .$$

Clearly, the lumped capacitance model will be appropriate in most cases and it will rarely be necessary to invoke a transmission line model.

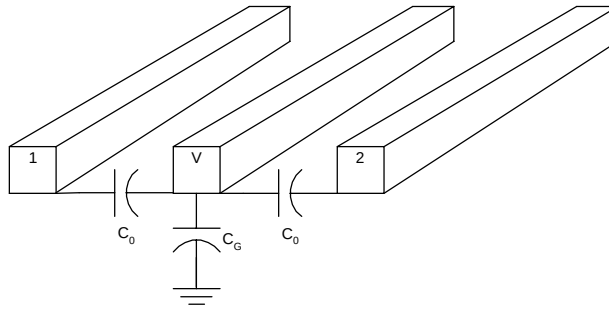


FIGURE 14.10
Parallel interconnects for the consideration of cross talk.

14.8 Special Problems in Interconnect Design

Special problems arise in the design of interconnect as a consequence of the parasitics. In data and address lines, the capacitive coupling between adjacent lines affects signal integrity and increases propagation delays. In the long polysilicon row lines used in memories, the rc delays are significant and represent an important design consideration. Because of their extent, clock distribution lines introduce significant delays along their length. Worse yet, unequal delays for various paths result in a situation called *clock skew*. In power distribution, the sizable currents can result in resistive drops that degrade the logic swings of downstream gates. This section will consider simple solutions to some of these challenges.

14.8.1 Cross Talk

In a multilevel metallization scheme, the interconnects exhibit parasitic capacitances to ground, to other interconnects on the same level (*interwire capacitance*), and to other interconnects on other levels (*interlevel capacitance*). The interwire parasitic capacitances result in the coupling of signals from the neighboring interconnects to the wire under consideration (the "victim"). This *cross talk* may compromise the integrity of signals and increase propagation delays. The worst manifestation of cross talk occurs in the case of data lines that are run side by side for a long distance, e.g., in a data bus, as shown in Figure 14.10.

The effective capacitance to ground for the victim wire is a function of the signals on the neighboring wires, as a consequence of the *Miller effect*. Suppose that the victim interconnect makes a low-to-high transition, with a logic swing equal to V_{DD}. If the adjacent lines are static and the interlevel contributions are ignored, then the effective capacitance of the victim line to ground is

TABLE 14.3

Cross Talk-Induced Miller Effect in Interconnect[a]

Victim	Neighbor 1	Neighbor 2	C_{eff}[a]
↑	↑	↑	C_G
↑	↑	↓	$C_G + 2C_0$
↑	↓	↑	$C_G + 2C_0$
↑	↓	↓	$C_G + 4C_0$
↓	↑	↑	$C_G + 4C_0$
↓	↑	↓	$C_G + 2C_0$
↓	↓	↑	$C_G + 2C_0$
↓	↓	↓	C_G

[a] C_{eff} is the effective capacitance to ground for the victim wire.

$$C_{eff} = \frac{\Delta Q}{\Delta V} = \frac{V_{DD}C_G + 2V_{DD}C_0}{V_{DD}} = C_G + 2C_0 . \tag{14.21}$$

If, however, the adjacent lines also make low-to-high transitions, no displacement current will flow in the interwire capacitances. For this case, the effective capacitance of the victim line is

$$C_{eff} = C_G . \tag{14.22}$$

Finally, if both neighboring wires make a high-to-low transition, then the voltage swings in the interwire capacitances are twice the logic swing on the lines. As a consequence, the effective capacitance to ground for the victim wire is increased to

$$C_{eff} = C_G + 4C_0 . \tag{14.23}$$

These various permutations are elaborated in Table 14.3. Here, the up and down arrows indicate low-to-high and high-to-low transitions, respectively. C_{eff} is the effective capacitance of the victim line to ground and C_G is the actual capacitance to ground for the victim line; the interwire capacitances to the neighboring wires are each assumed to be C_0.

As a result of cross talk, the effective capacitance to ground of the victim wire can be increased significantly, especially if the neighboring wires run parallel for any distance. Worse yet, this capacitance (and therefore the propagation delay of the driving gate) becomes a function of the signals on the other lines, in turn making performance predictions difficult.

There are several remedies to the cross-talk problem. One common approach is to run wires on adjacent levels in orthogonal directions as shown in Figure 14.11. This causes the interlevel capacitances to split into many small components, none of which is significant enough to cause cross-talk problems. Another

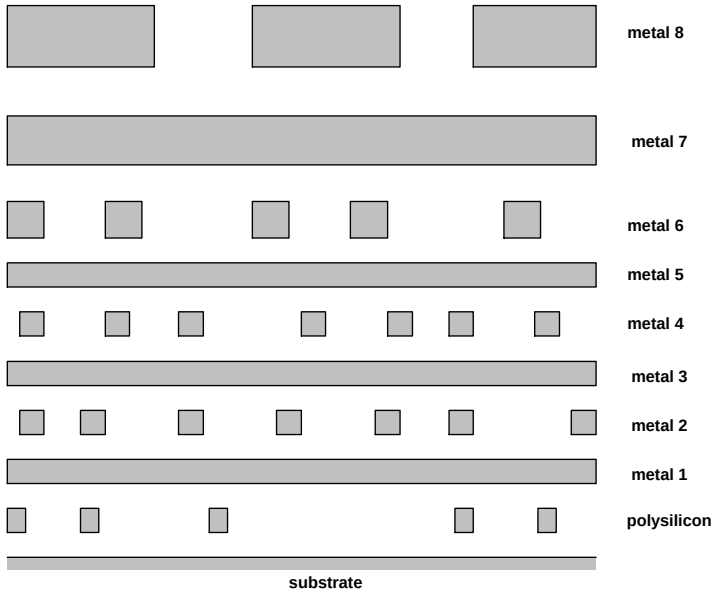


FIGURE 14.11
Orthogonal interconnects on adjacent levels minimize the cross talk between levels.

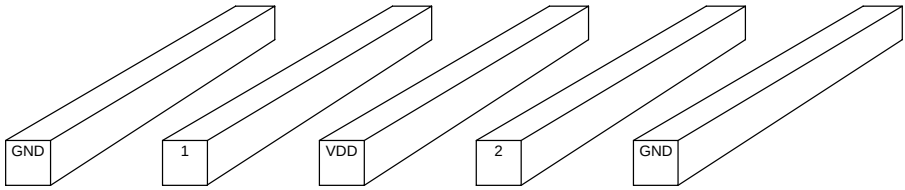


FIGURE 14.12
Power rails inserted between signal lines to reduce the effects of cross talk.

is to avoid long stretches of parallel lines by creative routing — an approach facilitated by computer routing tools. Yet another approach is the insertion of ground and V_{DD} lines between signal lines as illustrated in Figure 14.12. There is no Miller effect with respect to the power rails; therefore, although the total capacitance to ground is increased somewhat, it is predictable.

14.8.2 Polysilicon Interconnect

Polysilicon is highly resistive compared to metal, even when heavily doped n-type. Thus, long polysilicon interconnects introduce significant delays. However, in some situations, using long polysilicon runs increases the layout efficiency, e.g., the row lines in a digital memory. In such cases, it is necessary to reduce the interconnect rc delays by using low-resistivity straps (metal or silicide) or by using repeaters, or some combination of both.

Example 14.11

Estimate the delay associated with 100 μm of n-doped polysilicon interconnect, 0.25 μm wide and 0.25 μm thick, on a 0.5-μm thick layer of silicon dioxide ($\varepsilon_r = 3.9$).

Solution. The capacitance to the substrate per unit length is

$$c = \frac{\varepsilon_{DI} w}{t_{DI}} + \frac{\varepsilon_{DI}}{\log(1 + t_{DI}/h)}$$

$$= \frac{(3.9)(8.85\times10^{-14}\ \text{F/cm})(0.25\times10^{-4}\ \text{cm})}{0.5\times10^{-4}\ \text{cm}} + \frac{(3.9)(8.85\times10^{-14}\ \text{F/cm})}{\log(1+0.5/0.25)}$$

$$= 0.89\times10^{-12}\ \text{F/cm} = 0.89\ \text{pF/cm}$$

and the resistance per unit length is

$$r = \frac{\rho}{wh} = \frac{(0.01\ \Omega\text{cm})}{(0.25\times10^{-4}\ \text{cm})(0.25\times10^{-4}\ \text{cm})} = 16\ \text{M}\Omega/\text{cm}$$

The delay associated with 100 μm of this polysilicon interconnect is

$$t_P \approx \ln(2)\frac{rcl^2}{2} = \ln(2)\frac{(16\times10^6\ \Omega/\text{cm})(0.89\times10^{-12}\ \text{F/cm})(10^{-2}\ \text{cm})^2}{2} = 490\ \text{ps},$$

which exceeds the on-chip propagation delay for a modern CMOS logic circuit by more than an order of magnitude.

The long RC delays associated with polysilicon interconnect can be dealt with by strapping or the use of repeaters. Strapping involves implementing a parallel interconnect with lower resistance per unit length. One approach is to create a silicide layer on top of the polysilicon; this approach reduces the sheet resistivity by a factor of 1/100. Another approach is to run a conventional metal interconnect parallel to the polysilicon and make metal–polysilicon contacts at regular intervals along the path. The use of repeaters involves inserting buffers at periodic intervals as shown in Figure 14.13.

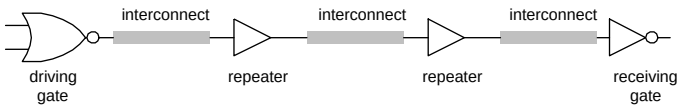


FIGURE 14.13
The use of repeaters to reduce interconnect delays.

Example 14.12

Estimate the reduction in the interconnect delay associated with 100 μm of n-doped polysilicon interconnect, 0.25 μm wide and 0.25 μm thick, on a 0.5-μm thick layer of silicon dioxide ($\varepsilon_r = 3.9$) by using nine repeaters.

Solution. The capacitance to the substrate per unit length is c = 0.89 pF/cm and the resistance per unit length is r = 16 MΩ/cm. Equal spacing of the nine repeaters (10 μm apart) will provide the maximum benefit. Assuming equal spacing, each 10-μm segment will contribute a delay equal to

$$t_P(\text{each segment}) \approx \ln(2)\frac{\left(16\times10^{6}\ \Omega/\text{cm}\right)\left(0.89\times10^{-12}\ \text{F/cm}\right)\left(10^{-3}\ \text{cm}\right)^2}{2}$$

$$= 4.9\ \text{ps}$$

If the repeater propagation delays can be neglected, the total delay is $t_p \approx 10t_p$ (each segment) = 49 ps, which is a reduction to 1/10 of the delay for the case without repeaters. Even if each repeater introduces a delay equal to the interconnect segments, there is a reduction to one fifth in the overall delay — a significant improvement.

14.8.3 Clock Distribution

Clock distribution presents special challenges because this signal must be delivered to every part of the die without appreciable phase differences. The individual interconnects in this distribution network will approach the size of the die, on the order of centimeters. Therefore, the individual wires in the clock distribution network will each introduce appreciable RC delays, even if they are made of copper embedded in a low-k dielectric. The elimination of *clock skew* (phase differences between different parts of the die) requires that these individual delays be made equal.

A common approach to this problem is to distribute the clock using an H tree, as shown in Figure 14.14. Here, the path lengths are equal, so, ideally, this makes all of the interconnect delays equal and eliminates the clock skew. However, the extent of such a clock distribution network may render its load capacitance in the order of nF, placing stringent demands on a single clock driver. Multiple drivers may therefore be used to share the load, as in Figure 14.14b.

14.8.4 Power Distribution

Distribution of the power rails (GND and V_{DD}) presents very different challenges compared to the case of clock distribution. Here, the large time-varying currents in the rails may cause a degradation of the local supply voltage and

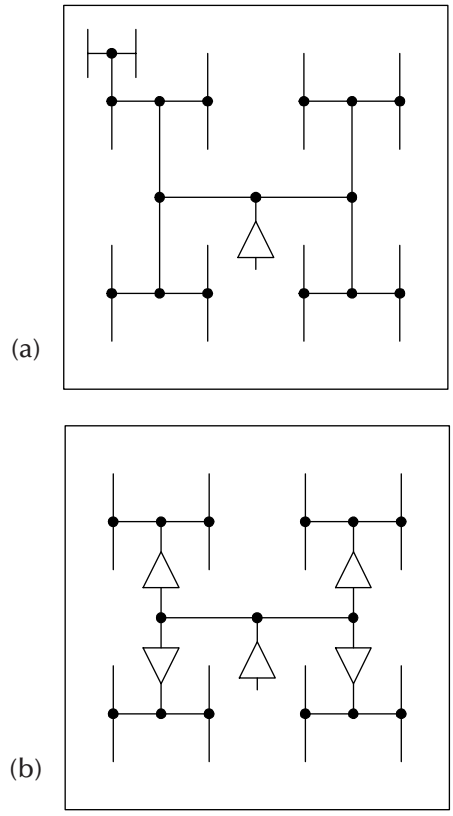


FIGURE 14.14
Clock distribution using an H-tree network with a single driver (a) and with multiple drivers (b).

therefore performance. Of special importance are the parasitic inductances and the associated voltage drops (the so-called Ldi/dt problem). However, the parasitic resistances also cause local degradation of the rail voltages.

Minimization of these problems requires massively parallel distribution networks for the power rails. This may be achieved in a single layer of interconnect using two arrays of parallel wires or in two levels by using a dual-plane approach, as shown in Figure 14.15. In ECL circuits, separate grounds are also used on the chip (the *clean* and *dirty* ground described in Chapter 6).

14.9 PSPICE Simulations

Simulations were performed using PSPICE 9.1, which can be downloaded free.[25]

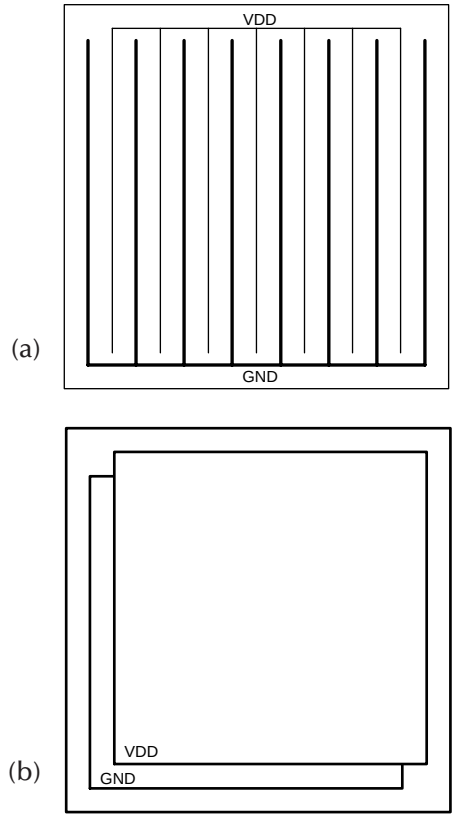


FIGURE 14.15
Power distribution networks: (a) single-level power grid and (b) dual plane power distribution.

14.9.1 Distributed RC Lines

A distributed RC line may be broken up into a convenient number of pieces for approximate SPICE analysis. Consider, for example, a polysilicon interconnect with $c = 0.89$ pF/cm and $r = 16$ MΩ/cm. For a 100 μm length of this interconnect, the expected propagation delay is

$$t_P \approx \ln(2)\frac{rcl^2}{2} = \ln(2)\frac{\left(16\times10^6\ \Omega/\text{cm}\right)\left(0.89\times10^{-12}\ \text{F/cm}\right)\left(10^{-2}\ \text{cm}\right)^2}{2} = 490\ \text{ps}.$$

The 100-μm interconnect may be modeled in SPICE using five RC sections, as shown in Figure 14.16.

As can be seen in Figure 14.17, the simulated propagation delays for 100 μm of polysilicon are 610 ps, or about 25% greater than the calculated delay. Therefore, using a finite number of sections provides reasonably accurate results while minimizing the computational overhead.

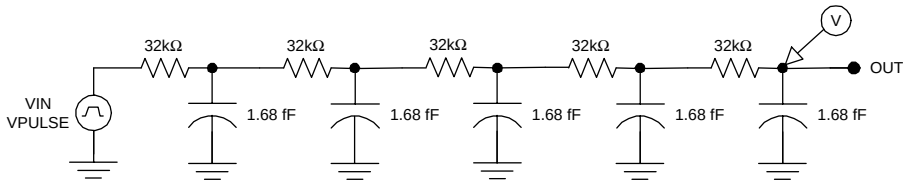


FIGURE 14.16
Circuit for PSPICE simulation of a 100-μm polysilicon interconnect with r = 16 MΩ/cm and c = 0.89 pF/cm, broken up into five identical RC sections.

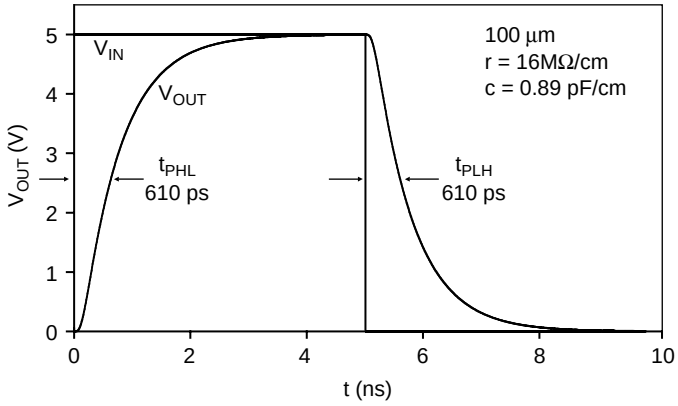


FIGURE 14.17
Simulated transient response for 100 μm of polysilicon with r = 16 MΩ/cm and c = 0.89 pF/cm, modeled by five identical RC sections with R = 32 kΩ and C = 1.68 fF.

14.9.2 Branched RC Lines

Consider branching interconnect as modeled in Figure 14.18. The propagation delays for nodes 1 and 5 with respect to the driving node may be estimated as

$$\begin{aligned} t_{P1} &\approx \ln(2)\sum_{k=1}^{N} C_k R_{1k} \\ &= \ln(2)\left[C_1R_1 + C_2R_1 + C_3R_1 + C_4R_1 + C_5R_1\right] \\ &= \ln(2)\left[\left(C_1 + C_2 + C_3 + C_4 + C_5\right)R_1\right] \\ &= \ln(2)\left[(540fF)(10\Omega)\right] = 3.7ps \end{aligned} \tag{14.24}$$

and

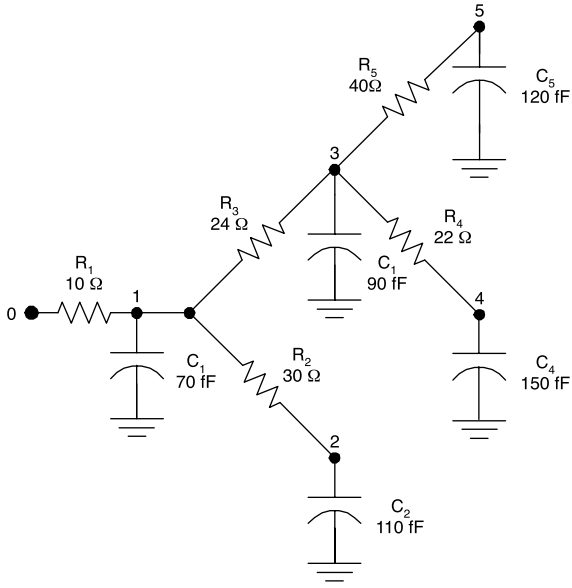


FIGURE 14.18
Branching interconnect modeled by a branching RC network.

$$t_{P5} \approx \ln(2)\sum_{k=1}^{N} C_k R_{1k}$$

$$= \ln(2)\left[C_1R_1 + C_2R_1 + C_3\left(R_1 + R_3\right) + C_4\left(R_1 + R_3\right) + C_5\left(R_1 + R_3 + R_5\right)\right].$$

$$= \ln(2)\left[\left(C_1 + C_2 + C_3 + C_4 + C_5\right)R_1 + \left(C_3 + C_4 + C_5\right)R_3 + C_5R_5\right]$$

$$= \ln(2)\left[\left(540fF\right)\left(10\Omega\right) + \left(360fF\right)\left(24\Omega\right) + \left(120fF\right)\left(40\Omega\right)\right] = 13.1ps$$

The delays may be determined in PSPICE using the circuit of Figure 14.19.

The results of the simulation appear in Figure 14.20. For node 5, the simulated propagation delay is 14.4 ps, or 10% more than the value predicted by the Elmore approximation. For node 1, however, the Elmore approximation overestimates the delay by more than a factor of four (3.7 vs. 0.8 ps). Therefore, caution must be used in applying the Elmore approximation to nodes near the driving node.

14.9.3 Transmission Lines

SPICE simulators provide models for ideal and lossy transmission lines. For example, in PSPICE the ideal transmission line is called "T" and the lossy transmission line part name is "Tlossy." For the ideal transmission line, the

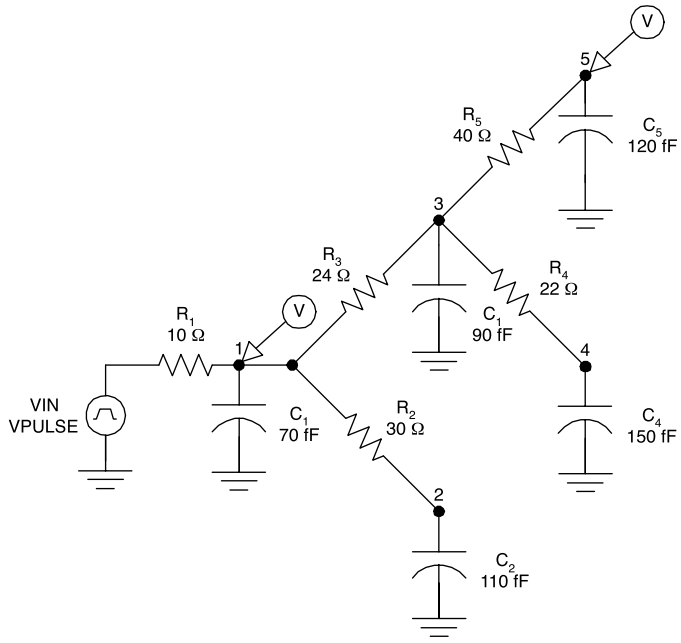


FIGURE 14.19
Circuit model for the PSPICE simulation of the transient response for a branching interconnect.

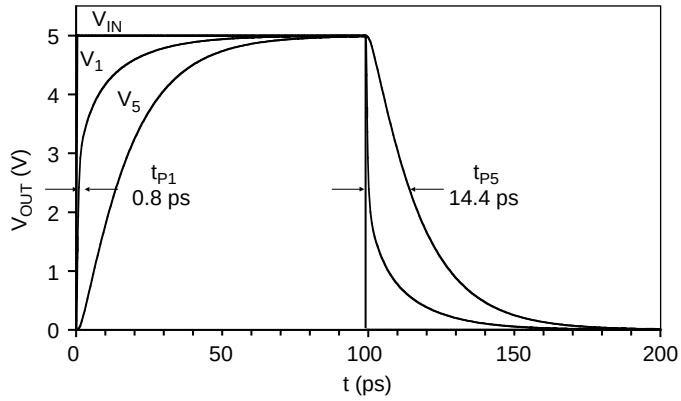


FIGURE 14.20
Simulated transient response for the branching interconnect.

only model parameters are the characteristic impedance ZO and the time of flight TD.

As an example, consider an ideal (lossless) transmission line connected to a pulse source and a terminating resistive load as shown in Figure 14.21. The model parameters for the transmission line are ZO = 50 Ω and TD = 66 ps. This time of flight corresponds to a 1 cm long transmission line embedded

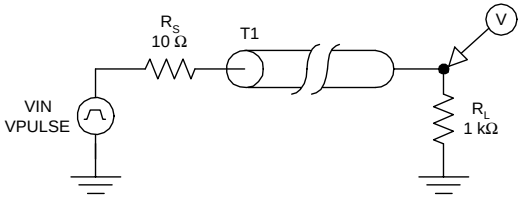


FIGURE 14.21
PSPICE circuit for transient simulation of a transmission line with a 1-kΩ termination.

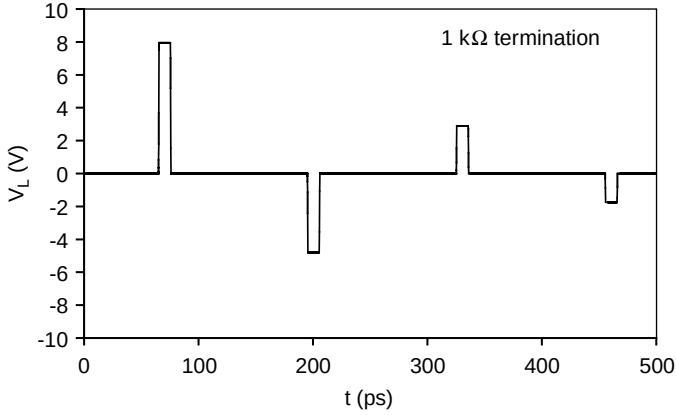


FIGURE 14.22
PSPICE transient simulation for a lossless transmission line with a 1-kΩ termination. For the transmission line, ZO = 50 Ω and TD = 66 ps. The pulse source has a 10-Ω impedance.

in silicon dioxide. The pulse source was set up with the following parameters: V1 = 0, V2 = 5 V, TR = 0, TF = 0, PW = 20 ps, and PER = 2000 ps. The source impedance is 10 Ω and the terminating load is 1 kΩ. For this configuration, the reflection coefficient at the load is 0.90 and the reflection coefficient at the source end is –0.67.

The transient simulation results for the transmission line of Figure 14.21 are shown in Figure 14.22. After 66 ps (the time of flight) a pulse is observed at the load. The amplitude of this pulse is the sum of the incident and reflected waves. After one more time-of-flight delay, the reflected wave is reflected from the source; therefore another pulse is observed after three times the time of flight, or 198 ps. Additional pulses are observed at a time interval equal to twice the time of flight. Each successive pulse is of opposite sign because of the negative reflection coefficient at the source.

If the transmission line is terminated with a latched 50-Ω load as in Figure 14.23, the behavior is very different, as shown in Figure 14.24; the result is a single pulse at the load, observed one time-of-flight delay after the pulse is launched from the source. The reflection coefficient at the matched load is zero. Therefore, reflected pulses do not travel back and forth in the line even though the source impedance is unmatched to the transmission line.

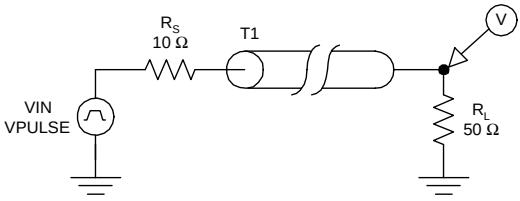


FIGURE 14.23
PSPICE circuit for transient simulation of a transmission line with a matched 50-Ω termination.

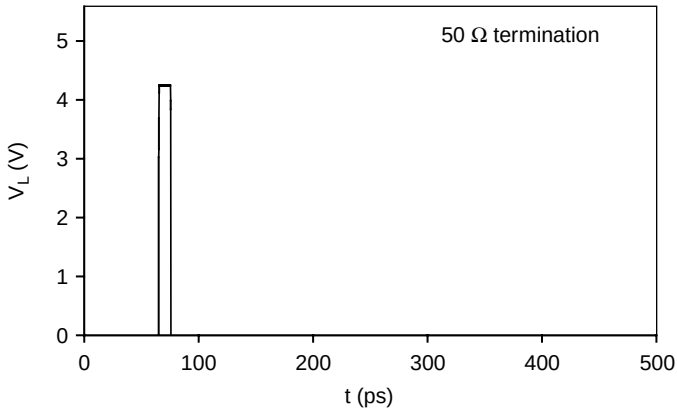


FIGURE 14.24
PSPICE transient simulation for a lossless transmission line with a matched 50-Ω termination. For the transmission line, ZO = 50 Ω and TD = 66 ps. The pulse source has a 10-Ω impedance.

An important conclusion to draw from this is that reflected pulses will be problematic in transmission lines that are terminated with unmatched impedances. This is true for real, lossy transmission lines as well. Therefore, interconnects or circuit board traces acting as transmission lines must be terminated properly in order to avoid reflections.

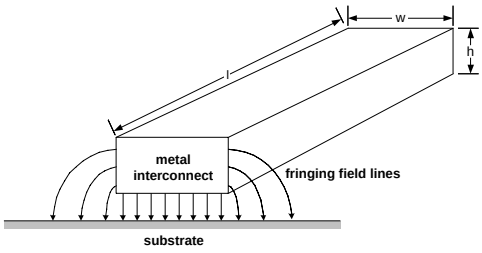
14.10 Summary

Because modern digital integrated circuits utilize about 1 km of interconnect per square centimeter of die area, interconnect has become an important consideration in their design, necessitating the use of low-resistance copper interconnects and low-k dielectrics in up to 10 levels.

Interconnects introduce parasitic capacitances, resistances, and inductances, which degrade circuit performance. Often these parasitics are quantized on a per-unit-length basis, using r, l, and c. The capacitance per unit length includes two components: the parallel plate capacitance and the fringing field capacitance. In practice, the fringing field component is dominant. In addition, the capacitances between wires are significant in a multilevel metallization scheme. The inductance per unit length may be estimated from the approximate value of c if the permeabilities and permittivities of the materials are known. However, the inductance is only important in very long interconnects or those providing power distribution. Short lengths of interconnect may be modeled using simple lumped capacitors.

INTERCONNECT QUICK REFERENCE

Parasitic Capacitance, Resistance, and Inductance of Interconnect



$$c = \frac{\varepsilon_{DI} w}{t_{DI}} + \frac{\varepsilon_{DI}}{\log(1 + t_{DI}/h)}$$

$$r = \frac{\rho}{wh}$$

$$l = \frac{\varepsilon_{DI}\mu_{DI}}{c}$$

Most digital systems combine gates from different circuit families; this necessitates interfacing. Often, the outputs of several gates are connected to a common node. This can result in wired logic. In other cases, transmission gates or tri-state gates must be used to avoid conflicts on common nodes such as those on data or address busses.

Lumped Capacitive Load

$$C_L = C_{interconnect} + C_{load}$$

Applicable if

$$l < \sqrt{\frac{t_{P,driver}}{rc\ln(2)/2}}$$

Distributed RC Line

$$t_P \approx \ln(2)\frac{rcl^2}{2}$$

Applicable if

$$l < \frac{t_{P,driver} c_0}{\sqrt{\varepsilon_r \mu_r}}$$

Transmission Line

$$t_{flight} = \frac{l\sqrt{\varepsilon_r \mu_r}}{c_0} \qquad Z_0 = \sqrt{\frac{l}{c}} \qquad \rho = \frac{R_L - Z_0}{R_L + Z_0}$$

Applicable if

$$l > \frac{t_{P,driver} c_0}{\sqrt{\varepsilon_r \mu_r}}$$

f = 10^{-15} p = 10^{-12} n = 10^{-9} μ = 10^{-6} m = 10^{-3} k = 10^{3} M = 10^{6} G = 10^{9}

Problems

P14.1. Estimate the low-frequency resistance per unit length for interconnects 0.25 μm wide and 1.0 μm in height made from

1. Copper
2. Aluminum
3. N-type polysilicon

P14.2. Estimate the 20-GHz skin depth and the resistance per unit length at this frequency for the interconnects considered in P14.1.

P14.3. Estimate the percentage increase in the resistance per unit length due to the skin effect for Cu interconnect with the following cross sections, at a frequency of 100 GHz:

1. 1 μm × 1μm
2. 5 μm × 1 μm

P14.4. Estimate the capacitance to ground per unit length for a copper interconnect 5.0 μm wide and 1.0 μm in height. The dielectric is silicon dioxide, 1.5 μm thick.

P14.5. Estimate the inductance per unit length for the interconnect of P14.4.

P14.6. Estimate the propagation delay for 100 μm of p-type polysilicon, 0.25 μm wide and 1.0 μm in height, using hand calculations. The dielectric is SiO_2, 1.5 μm thick. Repeat using SPICE by breaking the interconnect into 10 RC segments.

P14.7. Consider 200 μm of n-type polysilicon, 0.25 μm wide and 1.0 μm in height. What is the optimum number of repeaters that will result in the minimum total delay? Assume the repeaters are CMOS inverters, each with an unloaded propagation delay of 20 ps. The dielectric is SiO_2, 1.5 μm thick.

P14.8. For the branching interconnect modeled by a branching RC tree as depicted in Figure 14.25, estimate propagation delays from the source node to node 4 by hand calculations. Determine the propagation delay using SPICE and compare the results.

P14.9. Consider Cu level-one interconnect, 0.5 μm wide and 1.0 μm high, on 1.5 μm of carbon-doped silicon dioxide. Estimate the ranges of length for which the following models are appropriate: a) the lumped capacitance model, b) the distributed rc model, and c) the transmission line model. For the CMOS driving gate, $t_p = C_L$(1 ps/fF).

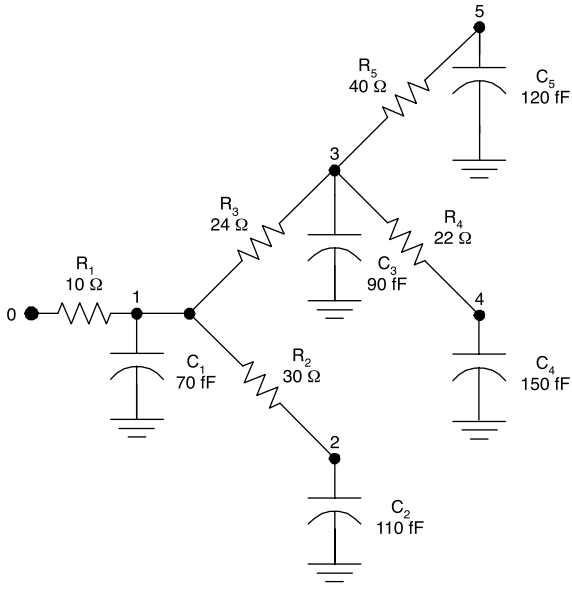


FIGURE 14.25
Branching interconnect modeled by an RC tree (P14.8).

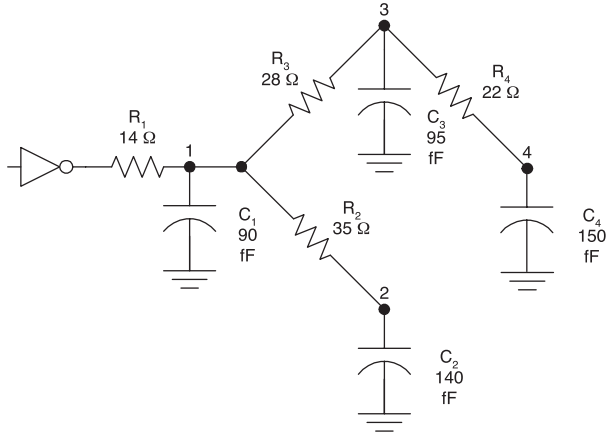


FIGURE 14.26
Branching interconnect modeled by an RC tree (P14.10).

P14.10. Consider the branching interconnect modeled by a branching RC tree as shown in Figure 14.26. Suppose the driving gate is a CMOS inverter with $K = 0.1$ mA/V^2, $V_{DD} = 1.5$ V, and $V_T = 0.3$ V. Can the resistances be neglected for determination of the propagation delays with respect to the input of the driving gate?

References

1. Meindl, J.D., Beyond Moore's law: the interconnect era, *Computing Sci. Eng.*, 5, 20, 2003.
2. Goel, A.K., Nanotechnology circuit design — the "interconnect problem," *Proc. 1st IEEE Conf. Nanotechnol.*, 123, 2001.
3. Davis, J.A., Venkatesan, R., Kaloyeros, A., Beylansky, M., Souri, S.J., Banerjee, K., Saraswat, K.C., Rahman, A., Reif, R., and Meindl, J.D., Interconnect limits on gigascale integration (GSI) in the 21st century, *Proc. IEEE*, 89, 305, 2001.
4. Mangaser, R. and Rose, K., Estimating interconnect performance for a new National Technology Roadmap for Semiconductors, *Proc. 1998 IEEE Interconnect Technol*. Conf., 253, 1998.
5. Toulouse, A., Bernard, D., Landrault, C., and Nouet, P., Efficient 3D modeling for extraction of interconnect capacitances in deep submicron dense layouts, *Proc. 1999 Design, Automation Test Eur. Conf.*, 576, 1999.
6. Chang, K.-J., Oh, S.-Y., and Lee, K., HIVE: an express and accurate interconnect capacitance extractor for submicron multilevel conductor systems, *Proc. 8th VLSI Multilevel Interconnection Conf.*, 359, 1991.
7. Choudhury, U. and Sangiovanni-Vincentelli, A., An analytical-model generator for interconnect capacitances, *Proc. 1991 IEEE Custom IC Conf.*, 8.6/1, 1991.
8. Wang, S.-Q., Spin-on dielectric films-a general overview, *Proc. 5th Int. Solid-State IC Technol. Conf.*, 961, 1998.
9. Taylor, K.J., Jeng, S.-P., Eissa, M., Gaynor, J., and Nguyen, H., Polymers for high performance interconnects, *Abstr. Booklet 1997 Euro. Workshop Mater. Adv. Metallization*, 59, 1997.
10. Ruelke, H., Streck, C., Hohage, J., Weiher–Telford, S., and Chretrien, O., Manufacturing implementation of low-k dielectrics for copper damascene technology, *Proc. 2002 IEEE Conf. Adv. Semiconductor Manuf.*, 356, 2002.
11. Lee, P.W., Lang, C.-I., Sugiarto, D., Xia, L.-Q., Gotuaco, M., and Yieh, E., Multigeneration CVD low κ films for 0.13 μm and beyond, *Proc. 6th Int. Conf. Solid-State IC Technol.*, 358, 2001.
12. Mosig, K., Jacobs, T., Kofron, P., Daniels, M., Brennan, K., Gonzales, A., Augur, R., Wetzel, J., Havemann, R., and Shiota, A., Single and dual damascene integration of a spin-on porous ultra low-k material, *Proc. 2001 IEEE Interconnect Technol. Conf.*, 292, 2001.
13. Kawakami, N., Fukumoto, Y., Kinoshita, T., Suzuki, K., and Inoue, K.-I., A super low-k (k = 1.1) silica aerogel film using supercritical drying technique, *Proc. IEEE Interconnect Technol. Conf.*, 143, 2000.
14. Naik, M., Parikh, S., Li, P., Educato, J., Cheung, D., Hashim, I., Hey, P., Jenq, S., Pan, T., Redeker, F., Rana, V., Tang, B., and Yost, D., Process integration of double level copper-low k (k = 2.8) interconnect, *1999 IEEE Int. Conf. Interconnect Technol.*, 181, 1999.
15. Lin–Hendel, C.G., Accurate interconnect modeling for high frequency LSI/VLSI circuits and systems, *Proc. 1990 IEEE Int. Conf. Computer Design: VLSI Computers Processors*, 434, 1990.
16. Ismail, Y.I., Friedman, E.G., and Neves, J.L., Equivalent Elmore delay for RLC trees, *IEEE Trans. Computer-Aided Design Integrated Circuits Syst.*, 19, 83, 2000.

17. Davis, J.A. and Meindl, J.D., Compact distributed RLC models for multilevel interconnect networks, *Dig. Tech. Papers 1999 Symp. VLSI Technol.*, 165, 1999.
18. Davis, J.A. and Meindl, J.D., Compact distributed RLC interconnect models — part II: coupled line transient expressions and peak crosstalk in multilevel networks, *IEEE Trans. Electron. Devices*, 47, 2078, 2000.
19. Venkatesan, R., Davis, J.A., and Meindl, J.D., Compact distributed RLC interconnect models — Part III: transients in single and coupled lines with capacitive load termination, *IEEE Trans. Electron. Devices*, 50, 1081, 2003.
20. Venkatesan, R., Davis, J.A., and Meindl. J.D., Compact distributed RLC interconnect models — Part IV: unified models for time delay, crosstalk, and repeater insertion, *IEEE Trans. Electron. Devices*, 50, 1094, 2003.
21. Elmore, E., The transient response of damped linear networks with particular regard to wideband amplifiers, *J. Appl. Phys.*, 55, 1948.
22. Yamakoshi, K. and Ino, M., Generalized Elmore delay expression for distributed RC tree networks, *Electron. Lett.*, 29, 617, 1993.
23. Abou-Seido, A.I., Nowak, B., and Chu, C., Fitted Elmore delay: a simple and accurate interconnect delay model, *Proc. 2002 IEEE Int. Conf. Computer Design: VLSI Computers Processors*, 422, 2002.
24. Adler, V. and Friedman, E., Uniform repeater insertion in RC trees, IEEE Trans. Circuits Sys. I: Fund. *Theory Appl.*, 47, 1, 2000.
25. www.cadence.com (Cadence).

15

Bistable Circuits

15.1 Introduction

Bistable circuits exhibit two stable states representing logic one and logic zero. These include latches and flip-flops,[1–4] which are useful in a number of applications that require the temporary retention of one or more bits. Some examples are counters, shift registers, and memories. Bistable circuits can also perform signal shaping functions, e.g., the Schmitt trigger, which exhibits hysteresis and is useful in this regard.[5,6]

The two requirements for realization of bistable operation are amplification (gain greater than unity) and positive feedback. A circuit meeting these requirements can be built using two cross-coupled inverters, as shown in Figure 15.1. There are two stable states for this circuit: state 0 is characterized by Q = 0 and $\overline{Q} = 1$ and state 1 is characterized by Q = 1 and $\overline{Q} = 0$. Either state is stable and will be maintained as long as the system power is on; therefore, this circuit can retain 1 bit of information.

The circuit of Figure 15.1 has limited usefulness because it has no input connections. Therefore, no way to set the circuit actively to state 0 or state 1 exists. Instead, the state is set randomly upon power-up. In contrast, practical bistable circuits provide inputs so that the state can be written as well as read.

The following sections consider latches, flip-flops, and Schmitt triggers. Latches are simple bistable circuits that have input connections to set their logic states; flip-flops are similar but are clocked. Schmitt triggers exhibit hysteresis; that is, the V_{IL} and V_{IH} are dependent on the output state. They are therefore useful for signal shaping applications. Digital memories are such an important class of bistable circuits that they will be covered separately in Chapter 16.

Example 15.1

Determine the stable states of the NMOS circuit illustrated in Figure 15.2.

Solution. The circuit exhibits two stable states. In state 0, M_{N0} is linear and M_{N1} is cut off. Therefore, for state 0, $V_{\overline{Q}} = 5$ V. V_Q may be found by equating the currents in M_{N0} and the left-hand pull-up resistor. Thus

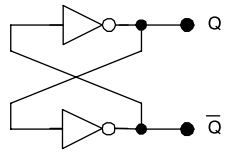


FIGURE 15.1
A bistable circuit constructed with cross-coupled inverters.

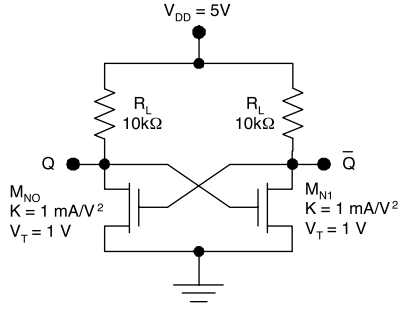


FIGURE 15.2
Cross-coupled NMOS inverters.

$$\frac{V_{DD} - V_Q}{R_L} = K\left[\left(V_{GS} - V_T\right)V_Q - \frac{V_Q^2}{2}\right].$$

Substituting in the circuit values,

$$\frac{5\text{ V} - V_Q}{10\text{ k}\Omega} = 1\text{ mA/V}^2\left[(4\text{ V})V_Q - \frac{V_Q^2}{2}\right]$$

or $(0.5\text{ mA/V}^2)V_Q^2 - (4.1\text{ mA/V})V_Q + 0.5\text{ mA} = 0$.

Applying the quadratic formula yields $V_Q = 4.1\text{ V} \pm 3.98\text{ V}$; therefore, in state 0, $V_Q = 0.12$ V.

Because of the symmetry of the circuit, the analysis of state 1 is similar. In state 1, M_{N0} is cut off and M_{N1} is linear, $V_Q = 0.12$ V and $V_{\overline{Q}} = 5$ V.

Does this circuit have a third stable state? Suppose that $V_Q = 0.80$ V. Figure 15.3 shows the circuit load line and the MOSFET characteristic corresponding to $V_{GS} = 1.80$ V. The two intersect at $V_{GS} = 1.80$ V; therefore, the solution is $V_{\overline{Q}} = 1.80$ V and, for the situation depicted, $V_Q = V_{\overline{Q}} = 1.80$ V. However, this is not a stable state. Because the loop gain of the circuit is greater than unity, any small noise voltage will cause the circuit to move to one of the two stable states.

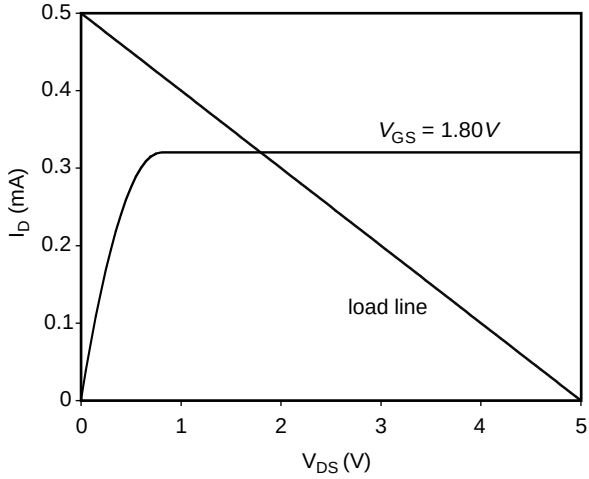


FIGURE 15.3
Load line and MOSFET characteristic for the case V_{GS} = 1.80 V.

15.2 RS Latch

The latch made from cross-coupled inverters does not allow active setting of the logic state, but this can be addressed by adding two MOSFETs, as shown in Figure 15.4. The left-hand MOSFET, M_{NIA}, allows active writing of the state 0. If the gate of M_{NIA} is brought to V_{DD}, the device will operate in the linear mode, bring V_Q low, and force the turn-off of M_{NOB}. This brings V_Q low and $V_{\bar{Q}}$ high; therefore, the left-hand input is called the "reset" input. Similarly, the application of V_{DD} at the right-hand input brings V_Q high and $V_{\bar{Q}}$ low and is therefore called the "set" input. The application of V_{DD} at both inputs should be avoided because in that case the final logic state is ambiguous. (The final logic state is determined by which high input lingers longest.)

This logic circuit is called a "reset set latch," or "RS latch" for short. Because it comprises cross-coupled NMOS NOR gates, it is also sometimes called a NOR latch. The logic representation and truth table for the NOR RS latch are shown in Figure 15.5. An RS latch may also be constructed using two cross-coupled NAND gates, as shown in Figure 15.6. This circuit differs from the NOR RS latch in one important regard. The set and reset inputs are active low, so that (0,0) is the input condition that must be avoided.

Usually, RS latches are not represented at the gate level, but instead use their own logic symbols as shown in Figure 15.7 and Figure 15.8.

In summary, a reset-set latch (RS latch) is a bistable circuit that can retain 1 bit of information. The RS latch may be constructed using cross-coupled NOR gates or cross-coupled NAND gates; the NOR version has active high

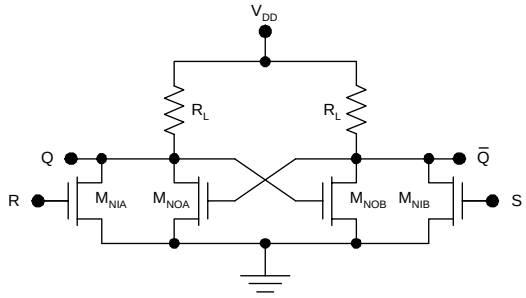
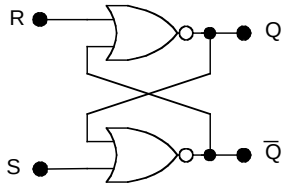
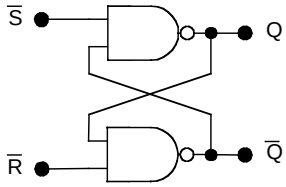


FIGURE 15.4
NMOS NOR RS latch.



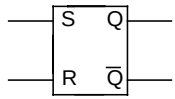
R	S	Q_{1+N}
0	0	Q_N
0	1	1
1	0	0
1	1	not used

FIGURE 15.5
NOR RS latch and truth table.



$\overline{R}$	$\overline{S}$	Q_{N+1}
0	0	not used
0	1	1
1	0	0
1	1	Q_N

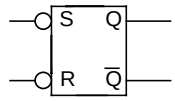
FIGURE 15.6
NAND RS latch and truth table.



R	S	Q_{N+1}
0	0	Q_N
0	1	1
1	0	0
1	1	not used

FIGURE 15.7
NOR RS latch symbol and truth table.

inputs. Bringing the S input to logic one "sets" the output Q to logic one and bringing the R input to logic one "resets" the output Q to logic zero. The input condition (S,R) = (1,1) must be avoided because of the associated ambiguity. In the NAND version of the RS latch, the inputs are active low. Bringing the S input to logic zero "sets" the output Q to logic one and



$\overline{R}$	$\overline{S}$	Q_{N+1}
0	0	not used
0	1	1
1	0	0
1	1	Q_N

FIGURE 15.8
NAND RS latch symbol and truth table.

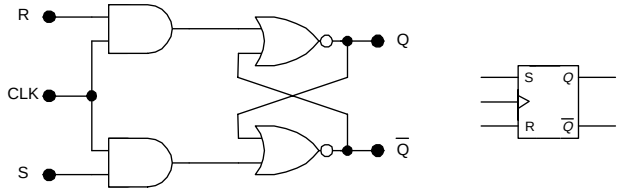


FIGURE 15.9
Clocked RS flip-flop.

bringing the R input to logic zero "resets" the output Q to logic zero. The input condition (S,R) = (0,0) must be avoided in the NAND RS latch. NOR latches are preferred in ECL, but NAND latches are preferred in TTL, CMOS, BiCMOS, and DCFL.

15.3 RS Flip-Flop

With RS latches, always an ambiguous input condition (R and S active) must be avoided. This situation can be alleviated somewhat by clocking the circuit, as shown in Figure 15.9. In the clocked circuit, the R and S inputs are inactive unless the clock signal is high. As long as R and S are established before the rising edge of the clock pulse and removed after the falling edge of the clock pulse, their exact timing is unimportant. By synchronizing the system with a clock in this way, it is possible to avoid inadvertent application of the forbidden input condition as a result of timing issues. As a matter of nomenclature, a clocked bistable circuit is called a flip-flop, whereas unclocked bistable circuits are called latches.

15.4 JK Flip-Flop

The ambiguous input condition of the RS latches and flip-flop can be avoided altogether by using feedback, as shown in Figure 15.10. The resulting device

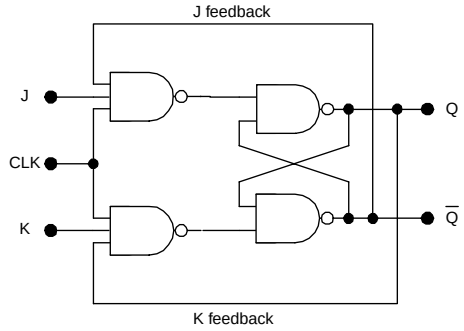
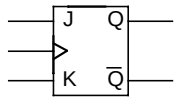


FIGURE 15.10
JK flip-flop.



J	K	Q_{N+1}
0	0	Q_N
0	1	1
1	0	0
1	1	$\overline{Q_N}$

FIGURE 15.11
JK flip-flop symbol and truth table.

is called a JK flip-flop, with the logic symbol and truth table as shown in Figure 15.11.

In the JK flip-flop, the J input is active only when $(CLK,\overline{Q}) = (1, 1)$ and the K input is active only when $(CLK,Q) = (1, 1)$; therefore, the two inputs will not be active simultaneously under static conditions. Thus, if (J,K) = (1,1), the output will "toggle" with each clock pulse. If the present output state of the flip-flop is Q_N, then the next state will be $Q_{N+1} = \overline{Q_N}$. However, the duration of the clock signal must be restricted to avoid the possibility of ambiguous operation. This can be seen in the timing diagram of Figure 15.12. The premise for this figure is that logic one is applied to J and to K, and a clock pulse of long duration is applied at $t = 0$. After two propagation delays, $\overline{Q}$ goes high, thus beginning the regeneration process; therefore, the clock duration must be at least two propagation delays. However, if the clock pulse persists for four propagation delays, both outputs will retoggle, which is undesirable because the final output state will depend on clock duration. For these reasons, reliable operation of the JK flip-flop requires a clock pulse with a duration greater than two propagation delays but less than four propagation delays:

$$2t_p < \frac{T}{2} < 4t_p \,. \tag{15.1}$$

Fortunately, the restriction on the duration of the clock pulse can be removed by using a master-slave design or edge triggering.

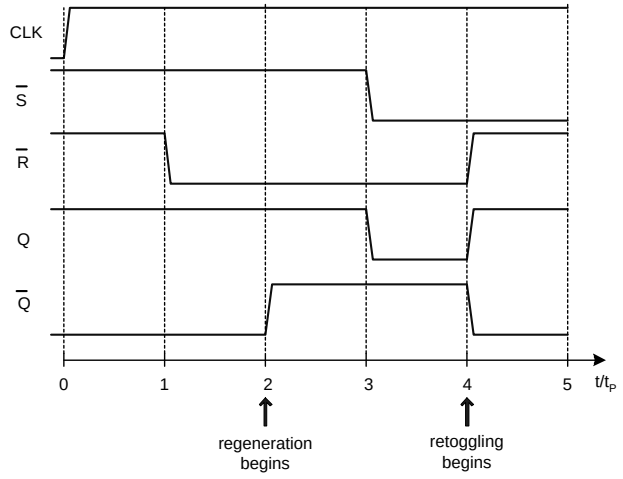


FIGURE 15.12
Timing diagram for the JK flip-flop.

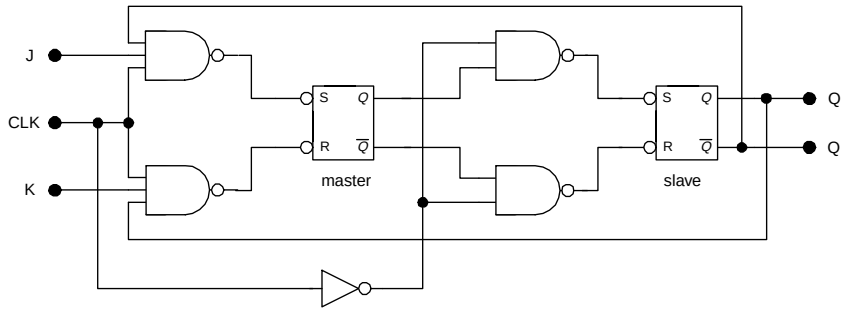


FIGURE 15.13
Master–slave JK flip-flop.

The master–slave design involves using two cascaded JK flip-flops as shown in Figure 15.13. Here, the left-hand ("master") flip-flop is only active while the clock signal is high. The right-hand ("slave") flip-flop uses an inverted version of this same clock and is only active when the external clock is low. The operation therefore proceeds as follows. When the clock goes high, the master switches, based on the values of J, K, and the output of the slave. When the clock goes low, the slave switches, based only on the output of the master. As a result and because the feedback connections are made from the output of the slave to the input of the master, output oscillations are not possible even with long-duration clock pulses.

The edge-triggered flip-flop exploits the differences in propagation delays for different paths within the circuits. One such design is shown in Figure 15.14. Upon the high-to-low transition of the clock, the signal CLK′ will go high after the propagation delay for gate 3 however, S′ and R′ will remain

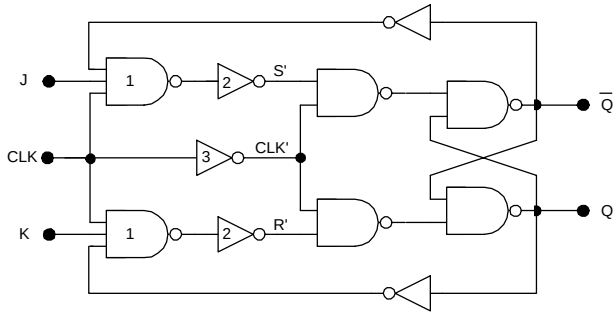


FIGURE 15.14
Edge-triggered JK flip-flop.

active for two propagation delays (the sum of the propagation delays for gates 1 and 2). Therefore, the three signals, S', R', and CLK', are simultaneously active for a time interval equal to

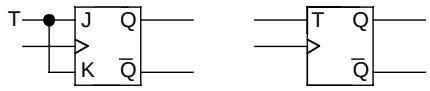
$$t_{active} = t_{P1} + t_{P2} - t_{P3}. \tag{15.2}$$

Moreover, because the flip-flop becomes active only for a short time after the high-to-low transition of the clock, this device is said to be "trailing edge-triggered." It is also quite possible to construct leading edge-triggered flip-flops, which become active for a short duration after the low-to-high transition of the clock. Although the master–slave and edge-triggering concepts have been explained separately for clarity, it is possible to incorporate both features in a single flip-flop. (This is nearly always done in practice.)

15.5 Other Flip-Flops

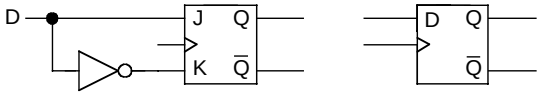
Several special-purpose flip-flops complement the JK type devices. Some of these can be realized by specially connecting the JK FF; two examples are the T ("toggle") flip-flop and the D ("data") flip-flop. The T flip-flop may be realized by connecting the J and K inputs of a JK FF as shown in Figure 15.15. The application of logic one to the T input causes the output to toggle (change state) upon each clock period. On the other hand, the application of logic zero to the T input causes the flip-flop to retain its current state indefinitely (as long as the power is maintained).

The D flip-flop may be realized by the configuration of a JK FF as shown in Figure 15.16. Here the D signal is applied directly to the J input, but the inverted version of D is applied to the K input. The application of logic one to the D input is characterized by J = 1 and K = 0, which causes the next state of the output to be 1. On the other hand, the application of logic zero



T	Q_{N+1}
0	Q_N
1	$\overline{Q_N}$

FIGURE 15.15
T flip-flop.



D	Q_{N+1}
0	0
1	1

FIGURE 15.16
D flip-flop.

to the D input is characterized by J = 0 and K = 1, thus causing the next state of the output to be 0. Therefore, the output of the D flip-flop always follows the bit of data applied at the input and the device therefore acts like a single-bit memory cell.

As with JK flip-flops, T and D flip-flops may utilize master–slave and edge-triggered concepts.

15.6 Schmitt Triggers

Schmitt triggers are specially constructed bistable circuits that exhibit hysteresis, i.e., V_{IL} and V_{IH} depend on the output state of the device. This property is useful in signal-shaping applications. In addition, Schmitt triggers display exceptional noise rejection capability because the presence of hysteresis makes it possible for the sum of the noise margins to exceed the supply voltage.

Schmitt triggers may be constructed using bipolar or field-effect transistors. In both cases, positive feedback and greater-than-unity loop gain are required, as with any bistable circuits. The achievement of hysteresis introduces another basic requirement: that a switching element introduce a state-dependent voltage between the input and ground. It is this circuit feature that introduces the hysteresis characteristic.

The hysteresis characteristic of a general Schmitt trigger inverter is shown in Figure 15.17. With the input at 0 V, the output voltage is V_{OH}. If the input voltage is increased, the output state will switch at the *upper trip voltage* V_U. In the output low state, the output voltage is V_{OL}; once the gate has switched to the output low state, an offset is introduced in the voltage transfer characteristic. If the input voltage is then swept from high to low, the output

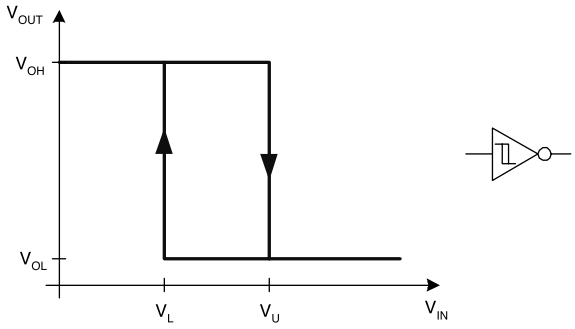


FIGURE 15.17
Schmitt trigger voltage transfer characteristic.

state will switch at the *lower trip voltage*, V_L. The difference between the trip voltages is called the *hysteresis voltage*, V_H, or sometimes simply *the hysteresis*:

$$V_H = V_U - V_L . \tag{15.3}$$

Schmitt triggers are distinguished from other logic gates by a hysteresis diagram imprinted within their symbols. An example is shown for the Schmitt trigger inverter in Figure 15.17. Schmitt triggers are superior in their noise rejection capabilities because they exhibit hysteresis. In nonhysteresis gates, the sum of the noise margins may not exceed the supply voltage; in Schmitt triggers, this restriction is removed. The noise margins for a Schmitt trigger are

$$V_{NMH} = V_{OH} - V_L \tag{15.4}$$

and

$$V_{NML} = V_U - V_{OL} . \tag{15.5}$$

Therefore, the sum of the noise margins for the Schmitt trigger is

$$V_{NML} + V_{NMH} = \left(V_{OH} - V_{OL}\right) + \left(V_U - V_L\right). \tag{15.6}$$

For a Schmitt trigger exhibiting rail-to-rail logic swing,

$$V_{NML} + V_{NMH} = V_{DD} + V_H . \tag{15.7}$$

With respect to noisy, slowly varying signals, hysteresis provides another benefit in addition to the increased noise margins. Suppose such a signal is

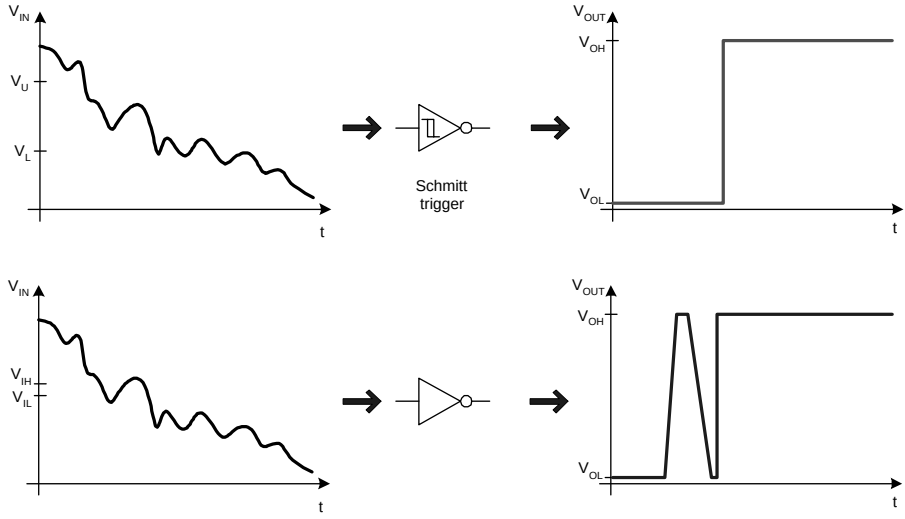


FIGURE 15.18
A noisy, slowly varying input signal is applied to a Schmitt trigger (top) and a conventional inverter (bottom).

applied to a conventional inverter and a Schmitt trigger inverter, as shown in Figure 15.18. The Schmitt trigger correctly interprets the waveform as a single high-to-low transition, whereas the nonhysteresis inverter misinterprets the input waveform. This difference is important if the result is to be used by a counter.

In addition to their ability to reject noise, Schmitt triggers are valued for their ability to sharpen slowly varying waveforms in the absence of noise. This is especially true in the case of CMOS, for which slowly varying waveforms give rise to excessive conduction and dissipation. The many Schmitt trigger circuits use different circuit topologies and perform different logic functions. No attempt will be made to catalog them here. Instead, two representative inverter designs will be examined in order to demonstrate the principles underlying these circuits.

15.6.1 Emitter-Coupled Schmitt Trigger

One version of the Schmitt trigger uses two bipolar transistors arranged with their emitters coupled together as shown in Figure 15.19. In this emitter-coupled Schmitt trigger, the output low state occurs with Q_{S2} saturated and Q_{S1} cutoff. If the input voltage is increased with the trigger in this state, the output state will abruptly switch when V_{IN} reaches the upper trip voltage. In this output high state, Q_{S2} is cut off and Q_{S1} is saturated. Then the input voltage must be lowered all the way to the lower trip voltage in order to cause the output state to switch back.

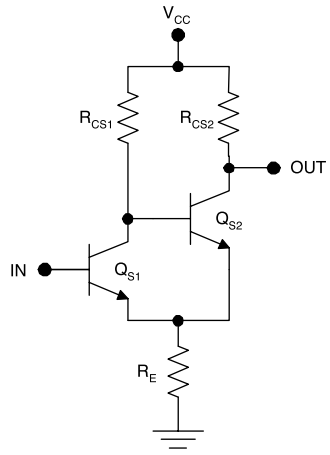


FIGURE 15.19
Emitter-coupled Schmitt trigger.

The trip voltages for the emitter-coupled Schmitt trigger may be determined as follows. In the output low state, with Q_{S2} saturated and Q_{S1} cut off,

$$I_{ES2} = I_{BS2} + I_{CS2}\,. \tag{15.8}$$

In terms of the voltage at the common emitter point, V_E,

$$\frac{V_E}{R_E} = \frac{V_{CC} - V_{BES} - V_E}{R_{CS1}} + \frac{V_{CC} - V_{CES} - V_E}{R_{CS2}}\,. \tag{15.9}$$

Solving for V_E,

$$V_E = \frac{R_{CS2}\left(V_{CC} - V_{BES}\right) + R_{CS1}\left(V_{CC} - V_{CES}\right)}{R_{CS1} + R_{CS2} + R_{CS1}R_{CS2}/R_E}\,. \tag{15.10}$$

The upper trip voltage is the value of the input voltage that causes Q_{S1} to turn on,

$$V_U = V_E + V_{BEA} = \frac{R_{CS2}\left(V_{CC} - V_{BES}\right) + R_{CS1}\left(V_{CC} - V_{CES}\right)}{R_{CS1} + R_{CS2} + R_{CS1}R_{CS2}/R_E} + V_{BEA}\,. \tag{15.11}$$

In order to determine the lower trip voltage, assume that the circuit is in the output high state, Q_{S2} is cut off, Q_{S1} is conducting, and the input voltage is being decreased. As V_{IN} approaches the lower trip voltage, Q_{S1} will operate in the forward active mode. Based on these assumptions, the common emitter voltage is

$$V_E = V_{\text{IN}} - V_{BEA}, \tag{15.12}$$

and the voltage at the base of Q_{S2} is

$$V_{BS2} = V_{CC} - R_{CS1}\left(\frac{V_{\text{IN}} - V_{BEA}}{R_E}\right)\left(\frac{\beta_F}{\beta_F + 1}\right). \tag{15.13}$$

The output state switches when Q_{S2} turns on, i.e., when

$$V_{BS2} - V_E = V_{BEA}; \tag{15.14}$$

thus, the lower trip voltage may be found from

$$V_{CC} - R_{CS1}\left(\frac{V_L - V_{BEA}}{R_E}\right)\left(\frac{\beta_F}{\beta_F + 1}\right) - V_L + V_{BEA} = V_{BEA}. \tag{15.15}$$

Solving, the lower trip voltage is

$$V_L = \frac{V_{CC} + V_{BEA}\left(\frac{R_{CS1}}{R_E}\right)\left(\frac{\beta_F}{\beta_F + 1}\right)}{\left(\frac{R_{CS1}}{R_E} + 1\right)}. \tag{15.16}$$

The hysteresis voltage is the difference between the trip voltages:

$$V_H = V_U - V_L \tag{15.17}$$

and therefore depends on the ratios of the resistor values.

Example 15.2

Determine the voltage transfer characteristic for the emitter-coupled Schmitt trigger depicted in Figure 15.20.

Solution. With Q_{S2} saturated,

$$V_E = \frac{R_{CS2}(V_{CC} - V_{BES}) + R_{CS1}(V_{CC} - V_{CES})}{R_{CS1} + R_{CS2} + R_{CS1}R_{CS2}/R_E}$$

$$= \frac{2.6\ \text{k}\Omega(4.2\ \text{V}) + 3.9\ \text{k}\Omega(4.9\ \text{V})}{3.9\ \text{k}\Omega + 2.6\ \text{k}\Omega + (3.9\ \text{k}\Omega)(2.6\ \text{k}\Omega)/1\ \text{k}\Omega} = 1.80\ \text{V}$$

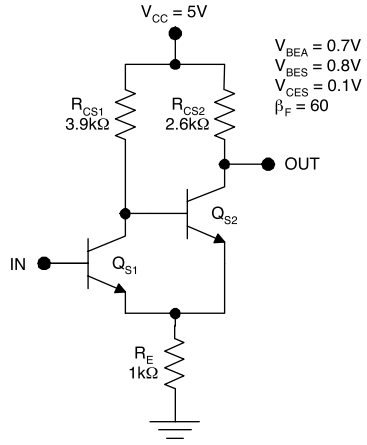


FIGURE 15.20
Example of emitter-coupled Schmitt trigger.

In this output low state, $V_{OL} = V_E + V_{CES} = 1.90$ V. The upper trip voltage is the value of V_{IN} that will cause Q_{S1} to turn on. It is $V_U = V_E + V_{BEA} = 2.50$ V. With Q_{S2} cut off, the output high voltage is $V_{OH} = V_{CC} = 5$ V and the lower trip voltage is

$$V_L = \frac{V_{CC} + V_{BEA}\left(\frac{R_{CS1}}{R_E}\right)\left(\frac{\beta_F}{\beta_F + 1}\right)}{\left(\frac{R_{CS1}}{R_E} + 1\right)} = \frac{5\text{ V} + 0.7\text{ V}\left(\frac{3.9\text{ k}\Omega}{1\text{ k}\Omega}\right)\left(\frac{60}{61}\right)}{\frac{3.9\text{ k}\Omega}{1\text{ k}\Omega} + 1} = 1.56\text{ V}$$

The resulting voltage transfer characteristic is shown in Figure 15.21.

Example 15.3

Determine the voltage transfer characteristic for the TTL NAND2 Schmitt trigger of Figure 15.22.

Solution. The trip voltages are lower by V_D compared to the trip voltages for the previous example. Thus, $V_L = 1.56$ V – 0.7 V = 0.86 V and $V_U = 2.50$ V – 0.7 V = 1.80 V. For the totem pole output, $V_{OL} = V_{CES} = 0.1$ V and $V_{OH} = V_{CC} - V_{BEA} - V_D = 3.6$ V. The resulting voltage transfer characteristic is plotted in Figure 15.23.

15.6.2 CMOS Schmitt Trigger

A Schmitt trigger can be realized in CMOS as shown in Figure 15.24. Here, the feedback transistors M_{PF} and M_{NF} introduce the hysteresis. The voltage transfer characteristic for the CMOS Schmitt trigger may be determined as

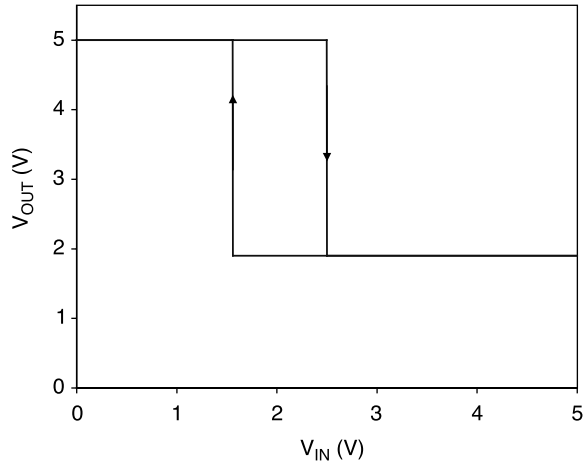


FIGURE 15.21
Voltage transfer characteristic for the example of emitter-coupled Schmitt trigger.

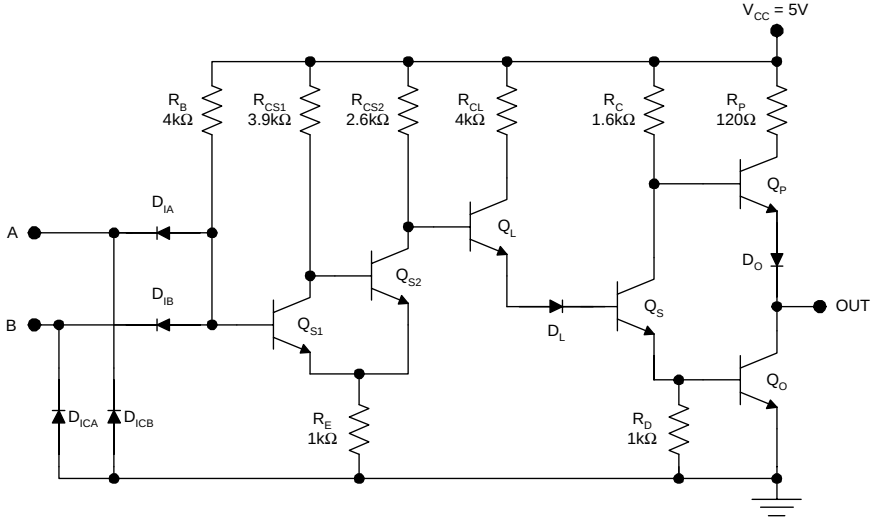


FIGURE 15.22
TTL NAND2 Schmitt trigger.

follows. With $V_{IN} = 0$, M_{NO}, M_{NI}, and the feedback transistors, M_{PF} and M_{NF}, are cut off while M_{PO} and M_{PI} are linear. Therefore,

$$V_{OH} = V_{DD}\,. \tag{15.18}$$

If V_{IN} is increased above V_T, then M_{NI} and M_{NF} are saturated. Together, these devices act as an NMOS inverter with a saturated enhancement type

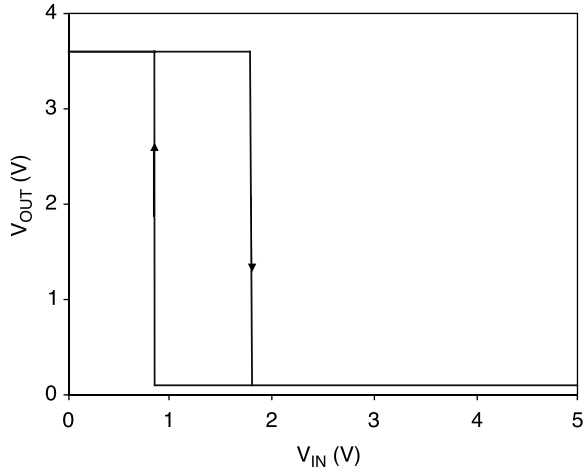


FIGURE 15.23
Voltage transfer characteristic for TTL Schmitt trigger.

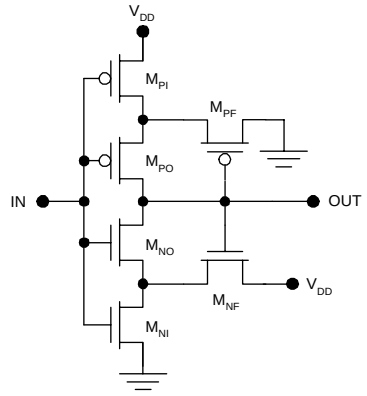


FIGURE 15.24
CMOS Schmitt trigger.

pull-up transistor. As long as the transistor M_{NO} does not conduct, the drain currents of M_{NI} and M_{NF} can be equated:

$$\frac{K_{NI}}{2}\left(V_{\text{IN}} - V_T\right)^2 = \frac{K_{NF}}{2}\left(V_{GSNF} - V_T\right)^2. \tag{15.19}$$

Solving, the gate-to-source voltage for the n-channel feedback transistor is

$$V_{GSNF} = \sqrt{\frac{K_{NI}}{K_{NF}}}\left(V_{\text{IN}} - V_T\right) + V_T\,; \tag{15.20}$$

therefore, the drain-to-source voltage for M_{NI} is

$$V_{DSNI} = V_{DD} - V_{GSNF} = V_{DD} - \sqrt{\frac{K_{NI}}{K_{NF}}}\left(V_{\text{IN}} - V_T\right) - V_T \,. \tag{15.21}$$

The upper trip voltage is the value of the input voltage that causes M_{NO} to turn on. In other words, at the trip voltage,

$$V_{GSNO} = V_{\text{IN}} - V_{DSNI} - V_T \,. \tag{15.22}$$

Solving, the upper trip voltage is found to be

$$V_U = \frac{V_{DD} + V_T\sqrt{\dfrac{K_{NI}}{K_{NF}}}}{1 + \sqrt{\dfrac{K_{NI}}{K_{NF}}}} \,. \tag{15.23}$$

To determine the lower trip voltage, suppose that the input voltage is decreased starting from V_{DD}. With $V_{\text{IN}} = V_{DD}$, M_{PO}, M_{PI}, and the feedback transistors, M_{PF} and M_{NF}, are cut off while M_{NO} and M_{NI} are linear. Therefore,

$$V_{OL} = 0 \,. \tag{15.24}$$

If V_{IN} is decreased below $V_{DD} - V_T$, M_{PI} and M_{PF} will operate in saturation. Together, these devices act like a *PMOS inverter** with a saturated enhancement type pull-up transistor. As long as the transistor M_{PO} does not conduct, the drain currents of M_{PI} and M_{PF} can be equated:

$$\frac{K_{PI}}{2}\left(V_{DD} - V_{\text{IN}} - \left|V_T\right|\right)^2 = \frac{K_{PF}}{2}\left(V_{GSPF} - \left|V_T\right|\right)^2 \,. \tag{15.25}$$

Solving, the absolute value of the (negative) gate-to-source voltage for the p-channel feedback transistor is

$$\left|V_{GSPF}\right| = \sqrt{\frac{K_{PI}}{K_{PF}}}\left(V_{DD} - V_{\text{IN}} - \left|V_T\right|\right) + \left|V_T\right| ; \tag{15.26}$$

* PMOS is a logic family that preceded NMOS; the circuits are built using only p-channel MOSFETs. The inverter utilizes one p-MOS pull-down transistor and one p-MOS pull-up transistor and the operation is qualitatively similar to that of NMOS. PMOS was commercially important in the early days of MOS technology because at that time ionic contaminants made it impossible to fabricate normally off n-channel transistors. Even PMOS microprocessors appeared on the market; however, as soon as the technology permitted, PMOS was displaced by superior NMOS circuitry.

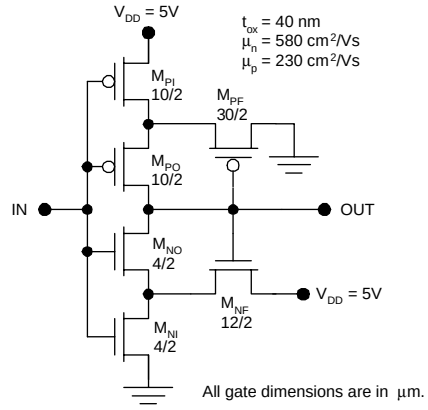


FIGURE 15.25
Example of CMOS Schmitt trigger.

therefore, the voltage at the source of M_{PO} with respect to ground is also

$$V_{SPO} = \sqrt{\frac{K_{PI}}{K_{PF}}}\left(V_{DD} - V_{\text{IN}} - |V_T|\right) + |V_T| . \tag{15.27}$$

The lower trip voltage is the value of the input voltage that causes M_{PO} to turn on. In other words, at the lower trip voltage,

$$V_{GSPO} = V_{\text{IN}} - V_{SPO} = -V_T . \tag{15.28}$$

Solving, the lower trip voltage is found to be

$$V_L = \frac{\left(V_{DD} - V_T\right)\sqrt{\dfrac{K_{PI}}{K_{PF}}}}{1 + \sqrt{\dfrac{K_{PI}}{K_{PF}}}} . \tag{15.29}$$

Example 15.4

Determine the voltage transfer characteristic for the CMOS Schmitt trigger of Figure 15.25.

Solution. With the gate dimensions given, the device transconductance parameters are related as follows: $K_{PI} = K_{PO} = K_{NO} = K_{NI}$, $K_{PF} = 3K_{PI}$, and $K_{NF} = 3K_{NI}$. The absolute values need not be known for the determination of the *voltage transfer characteristic*. The output voltage levels are $V_{OL} = 0$ and $V_{OH} =$ 5 V. The trip voltages are

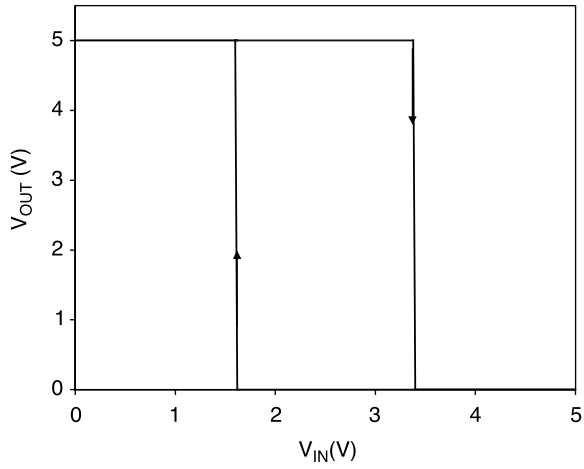


FIGURE 15.26
Example voltage transfer characteristic for CMOS Schmitt trigger.

$$V_U = \frac{V_{DD} + V_T\sqrt{\frac{K_{NI}}{K_{NF}}}}{1+\sqrt{\frac{K_{NI}}{K_{NF}}}} = \frac{5\text{ V} + 0.6\text{ V}\sqrt{1/3}}{1+\sqrt{1/3}} = 3.39\text{ V}$$

and

$$V_L = \frac{\left(V_{DD} - V_T\right)\sqrt{\frac{K_{PI}}{K_{PF}}}}{1+\sqrt{\frac{K_{PI}}{K_{PF}}}} = \frac{(5\text{ V} - 0.6\text{ V})\sqrt{1/3}}{1+\sqrt{1/3}} = 1.61\text{ V}\,.$$

Therefore, the circuit exhibits hysteresis of 1.78 V as shown in the voltage transfer characteristic of Figure 15.26.

15.7 PSPICE Simulations

Simulations were performed using PSPICE 9.1, which can be downloaded free.[7] The model parameters for bipolar devices are given in Table 15.1 and Table 15.2 and the MOSFET model parameters are provided in Table 15.3 and Table 15.4.

TABLE 15.1

BJT SPICE Parameters

Parameter	Value	Units
IS	2.0f	A
BF	70	—
NF	1	—
BR	0.5	—
NR	1	—
CJE	0.3p	F
VJE	0.8	V
MJE	0.333	—
TF	0.2n	s
CJC	0.15p	F
VJC	0.75	V
MJC	0.5	—
TR	10n	s

TABLE 15.2

Diode SPICE Parameters

Parameter	Value	Units
IS	2.0f	A
N	1	—
TT	0.2n	s
CJO	0.3p	F
VJ	0.8	V
M	0.5	—

TABLE 15.3

n-MOSFET SPICE Parameters

Parameter	Value	Units
VTO	0.5	V
KP	0.2m	A/V^2
LAMBDA	0.05	—
CGSO	1.15n	F/m
CGDO	0.58n	F/m
VMAX	100k	m/s

TABLE 15.4

p-MOSFET SPICE Parameters

Parameter	Value	Units
VTO	–0.5	V
KP	0.2m	A/V^2
LAMBDA	0.05	—
CGSO	1.15n	F/m
CGDO	0.58n	F/m
VMAX	80k	m/s

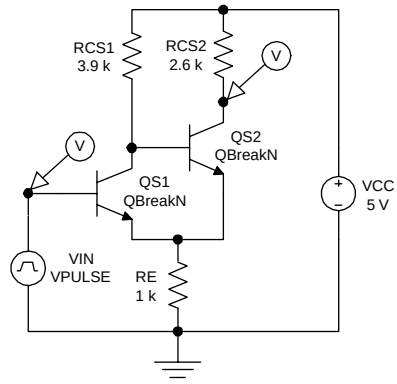


FIGURE 15.27
Emitter-coupled Schmitt trigger circuit used for simulation of the transient response.

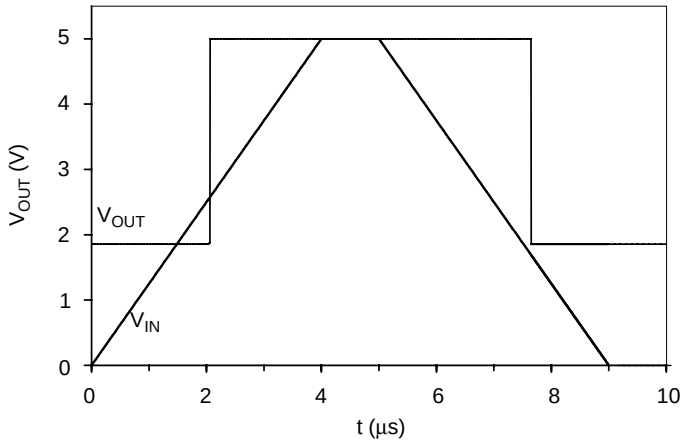


FIGURE 15.28
Simulated transient response for the emitter-coupled Schmitt trigger.

15.7.1 Emitter-Coupled Schmitt Trigger

Both trip voltages for a Schmitt trigger may be determined using a single transient simulation. This was done using the emitter-coupled Schmitt trigger circuit of Figure 15.27 and a slowly varying pulse input; the results of the transient simulation are shown in Figure 15.28. The trip voltages are the values of V_{IN} corresponding to the abrupt transitions in the output: V_L = 1.67 V and V_U = 2.57 V. These are close to the hand-calculated values of 1.56 V and 2.50 V, respectively.

The trip voltages may also be determined from the VTC. However, this requires two DC sweeps (V_{IN} must be swept from zero to V_{CC} and also from V_{CC} to zero.) This was done for the emitter-coupled Schmitt trigger circuit of Figure 15.29 and the results of both DC sweeps have been combined in Figure 15.30. The trip voltages thus determined are the same as those obtained from the transient analysis.

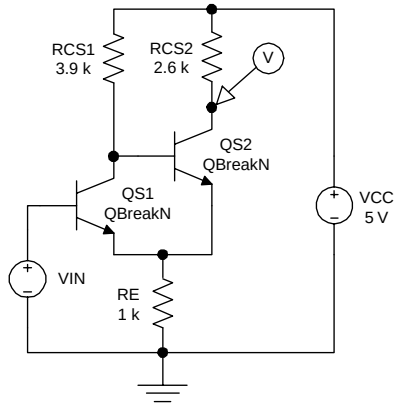


FIGURE 15.29
Emitter-coupled Schmitt trigger circuit used for simulation of the VTC.

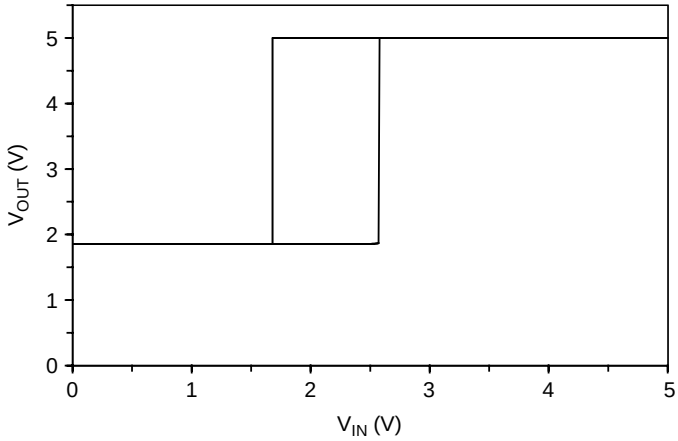


FIGURE 15.30
Simulated VTC for the emitter-coupled Schmitt trigger.

15.7.2 TTL Schmitt Trigger

The voltage transfer characteristic for the TTL Schmitt trigger inverter of Figure 15.31 was determined using two DC sweeps. As seen in the results of Figure 15.32, the presence of the input diodes shifts both trip voltages in the negative direction by V_D.

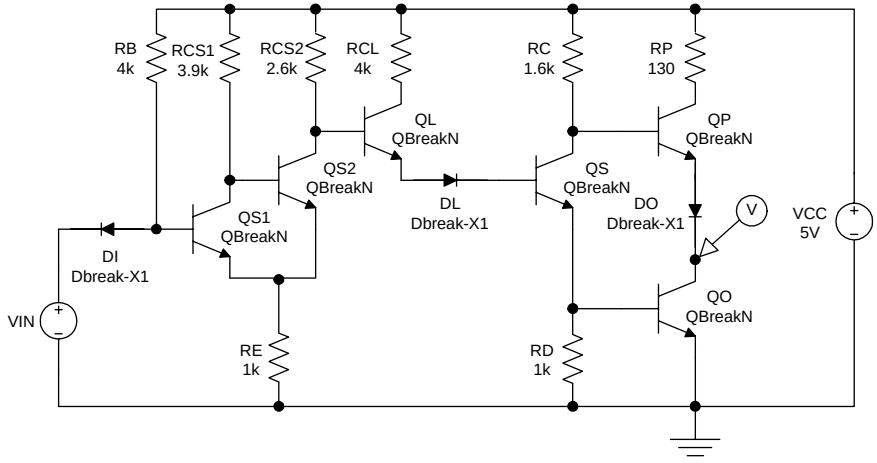


FIGURE 15.31
TTL Schmitt trigger circuit for simulation of the VTC.

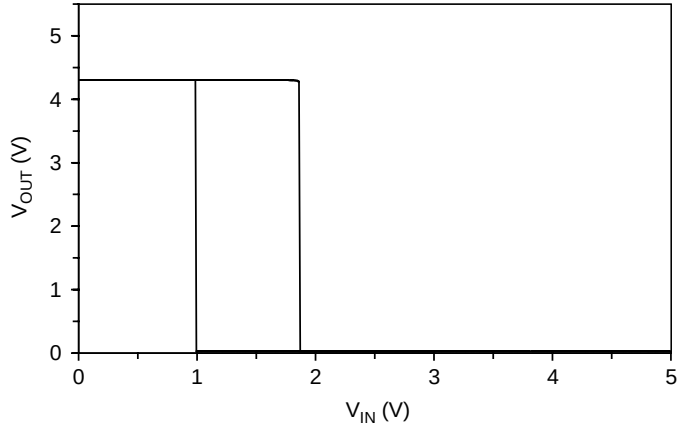


FIGURE 15.32
Simulated VTC for the TTL Schmitt trigger.

15.7.3 CMOS Schmitt Trigger

Two DC sweeps were used to determine the VTC for the CMOS Schmitt trigger of Figure 15.33. The results have been combined in Figure 15.34 and reveal trip voltages of 1.5 and 3.5 V.

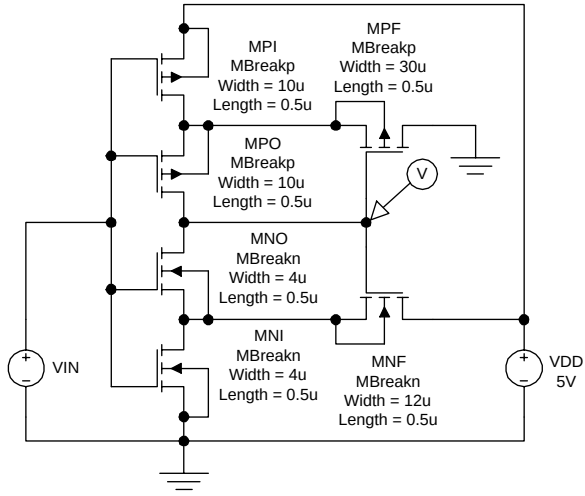


FIGURE 15.33
CMOS Schmitt trigger circuit for simulation of the VTC.

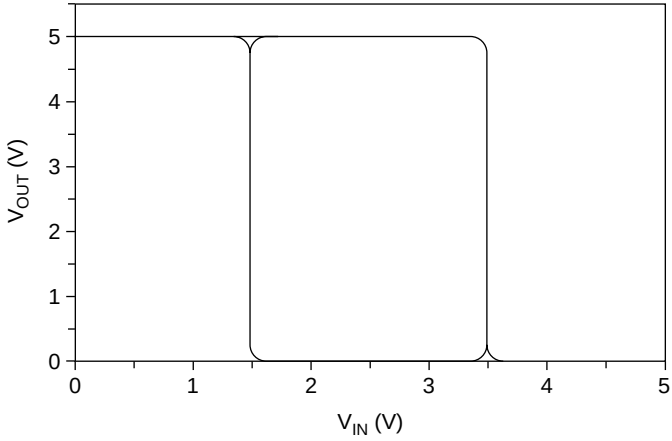


FIGURE 15.34
Simulated VTC for the CMOS Schmitt trigger.

15.8 Summary

Bistable circuits exhibit two stable output states and are therefore capable of retaining single bits of data. Applications of bistable circuits include counters, shift registers, and wave shaping. The three broad classes of bistable circuits are latches, flip-flops, and Schmitt triggers. Latches are

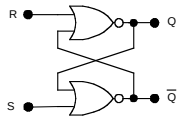
unclocked bistable circuits, whereas flip-flops are clocked. Schmitt triggers are specially constructed bistable circuits that exhibit hysteresis; they are useful in signal-shaping applications.

BISTABLE CIRCUITS QUICK REFERENCE

Bistable Circuits

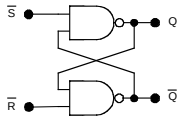
Bistable circuits exhibit two stable states and can be used to store one bit of information. The simplest bistable circuit is the latch, which is constructed from cross coupled NOR or NAND gates. Flip flops are clocked bistable circuits. Schmitt triggers exhibit hysteresis.

Latches and Flip Flops



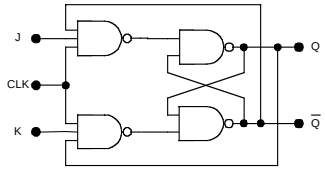
R	S	Q_{N+1}
0	0	Q_N
0	1	1
1	0	0
1	1	not used

NOR RS Latch



$\overline{R}$	$\overline{S}$	Q_{N+1}
0	0	not used
0	1	1
1	0	0
1	1	Q_N

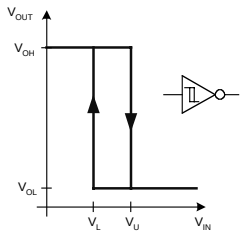
NAND RS Latch



J	K	Q_{N+1}
0	0	Q_N
0	1	0
1	0	1
1	1	$\overline{Q_N}$

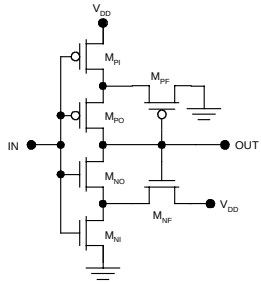
JK Flip Flop

Schmitt Triggers



$$V_H = V_U - V_L$$

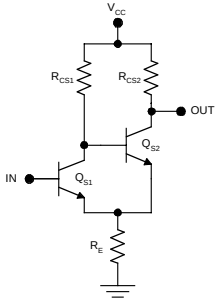
CMOS Schmitt Trigger



$$V_L = \frac{(V_{DD} - V_T)\sqrt{\frac{K_{PI}}{K_{PF}}}}{1 + \sqrt{\frac{K_{PI}}{K_{PF}}}}$$

$$V_U = \frac{V_{DD} + V_T\sqrt{\frac{K_{NI}}{K_{NF}}}}{1 + \sqrt{\frac{K_{NI}}{K_{NF}}}}$$

Emitter-Coupled Schmitt Trigger



$$V_L = \frac{V_{CC} + V_{BEA}\left(\frac{R_{CS1}}{R_E}\right)\left(\frac{\beta_F}{\beta_F + 1}\right)}{\left(\frac{R_{CS1}}{R_E} + 1\right)}$$

$$V_U = \frac{R_{CS2}(V_{CC} - V_{BES}) + R_{CS1}(V_{CC} - V_{CES})}{R_{CS1} + R_{CS2} + R_{CS1}R_{CS2}/R_E} + V_{BEA}$$

f = 10^{-15} p = 10^{-12} n = 10^{-9} μ = 10^{-6} m = 10^{-3} k = 10^{3} M = 10^{6} G = 10^{9}

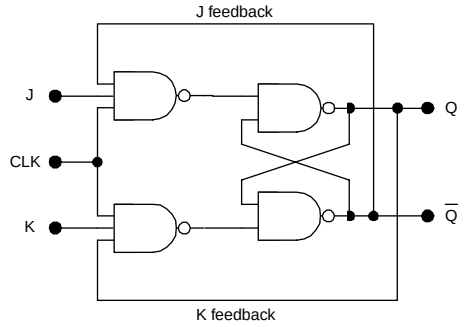


FIGURE 15.35
A simple JK flip-flop (L15.1).

Laboratory Exercises

L15.1. Using 74HCxx series CMOS NAND, NOR, and NOT gates, construct a simple JK flip-flop as shown in Figure 15.35. It is recommended that the flip-flop be constructed on a printed circuit board. This can be done using an Internet supplier of printed circuit boards by one of two means. The first is to download the manufacturer's CAD software for the design of the circuit board. The second is to create a Gerber file (if your version of SPICE supports this) and then submit the circuit board design electronically.

1. Using a 1-kHz clock, verify the operation of the flip-flop for all but the (J,K) = (1,1) condition.
2. Apply the (J,K) = (1,1) condition and determine the range of the clock period that results in reliable toggling operation of the flip-flop. Is this result as expected based on the propagation delays for 74HCxx circuits?
3. Apply the (J,K) = (1,1) condition and determine the range of the clock period that results in unwanted oscillations.

L15.2. Using commercially available devices, design a TTL Schmitt trigger that will exhibit greater than 1.5-V hysteresis using a 5-V supply.

1. Obtain the data sheets and SPICE models (if available) for the transistors and diodes from the manufacturer's Web pages.
2. Using the manufacturer's SPICE models (or custom models created using the information from the data sheets), verify the performance of the Schmitt trigger using SPICE.
3. Build the Schmitt trigger and measure the voltage transfer characteristic using the x–y feature of an oscilloscope or virtual instrument and thus determine the trip voltages and hysteresis voltage.

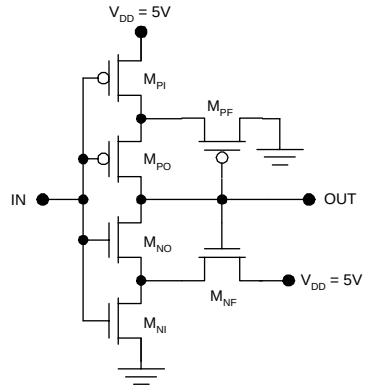


FIGURE 15.36
CMOS Schmitt trigger (L15.3).

L15.3. Using commercially available enhancement-type MOSFETs, design a CMOS Schmitt trigger that will exhibit greater than 2.3-V hysteresis using a 5-V supply and the circuit diagram of Figure 15.36. (MOSFETs may be paralleled to increase the W/L ratio effectively.)

1. Obtain the data sheets and SPICE models (if available) for the MOSFETs from the manufacturer's Web pages.
2. Using the manufacturer's SPICE models (or custom models created using the information from the data sheets), verify the performance of the Schmitt trigger using SPICE.
3. Build the Schmitt trigger and measure the voltage transfer characteristic using the x–y feature of an oscilloscope or virtual instrument and thus determine the trip voltages and hysteresis voltage.

L15.4. Build the circuit shown in Figure 15.37 to compare a Schmitt trigger to a conventional inverter.

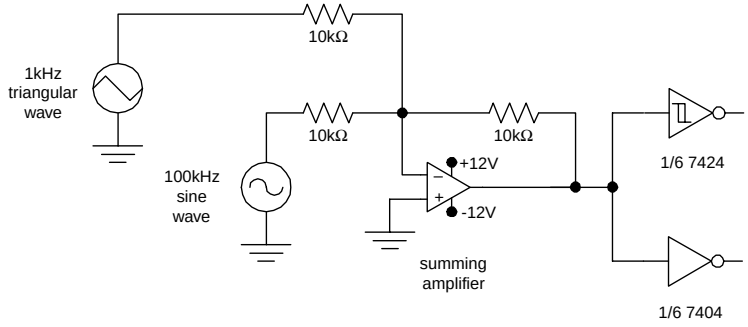


FIGURE 15.37
Circuit for testing Schmitt triggers (L15.4).

1. Use a 5-V supply for the Schmitt trigger. Set the peak-to-peak amplitude of the triangular wave to 5 V and the DC offset to 2.5 V. This simulates a slowly varying digital signal. Now increase the peak-to-peak amplitude of the sine wave source (which simulates noise) until bit errors are observed at the output. Determine this peak-to-peak value; is it what would be expected?
2. Use a 5-V supply for the conventional inverter. Set the peak-to-peak amplitude of the triangular wave to 5 V and the DC offset to 2.5 V. As in (a), increase the peak-to-peak amplitude of the sine wave source until bit errors are observed at the output. Determine this peak-to-peak value; is it what would be expected? How do the two circuits compare in their ability to reject noise?

Problems

P15.1. For the NMOS NOR latch shown in Figure 15.38, use SPICE to verify the logic function of the latch. (Apply pulse inputs to S and R. Make the frequency of the S input twice that of the R input and observe the output Q.) t_{ox} = 13 nm, μ_n = 580 cm^2/Vs, μ_p = 230 cm^2/Vs, and L = 1 μm. V_T = 0.6 V. R_L = 50 KΩ.

P15.2. Consider a simple JK flip-flop constructed using 74HCxx series CMOS gates as shown in Figure 15.39. (When using SPICE, construct the circuit at the MOSFET level unless your version of SPICE supports gate-level simulations.)

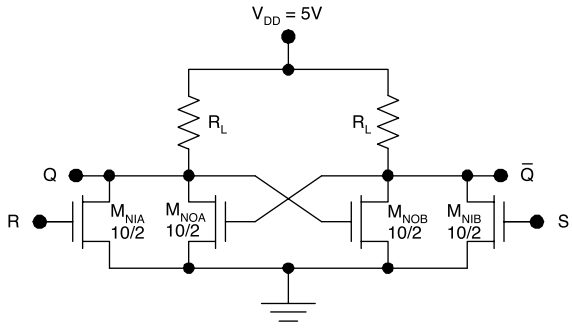


FIGURE 15.38
NMOS NOR latch (P15.1).

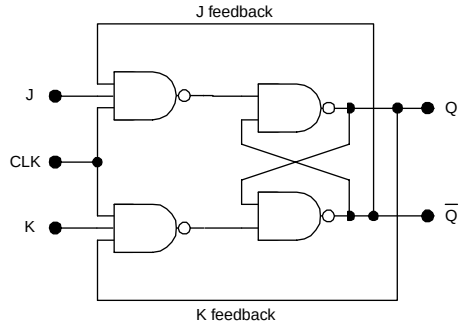


FIGURE 15.39
A simple JK flip-flop constructed using 74HCxx series CMOS gates (P15.2).

1. Using SPICE and a 1-kHz clock, verify the operation of the flip-flop for all but the (J,K) = (1,1) condition.
2. Using SPICE, apply the (J,K) = (1,1) condition and determine the range of the clock period that results in reliable toggling operation of the flip-flop. Is this result as expected based on the propagation delays for 74HCxx circuits?
3. Apply the (J,K) = (1,1) condition and determine the range of the clock period that results in unwanted oscillations.

P15.3. Consider the master–slave JK flip-flop constructed using 74HCxx series CMOS gates as illustrated in Figure 15.40. (When using SPICE, construct the circuit at the gate level.) Using SPICE and a 1-kHz clock, verify the operation of the flip-flop for all input conditions and show that oscillations are not observed for the (J,K) = (1,1) condition.

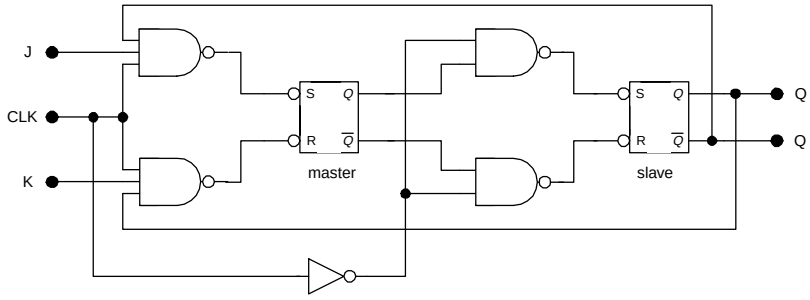


FIGURE 15.40
Master–slave JK flip-flop constructed using 74HCxx series CMOS gates (P15.3).

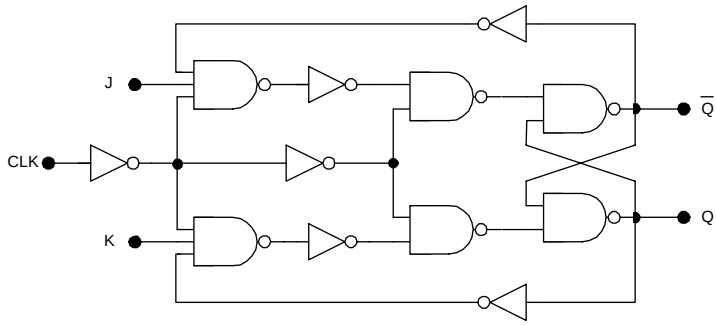


FIGURE 15.41
Edge-triggered JK flip-flop (P15.4).

P15.4. Consider the edge-triggered JK flip-flop of Figure 15.41, constructed using 74HCxx CMOS gates.

1. Is this flip-flop leading edge-triggered or trailing edge-triggered?
2. Verify this result by using SPICE to create the timing waveforms with a 1-kHz clock.

P15.5. Consider the emitter-coupled Schmitt trigger depicted in Figure 15.42.

1. Determine the trip voltages by hand calculations.
2. Determine the trip voltages using SPICE; compare the SPICE and hand-calculated results.

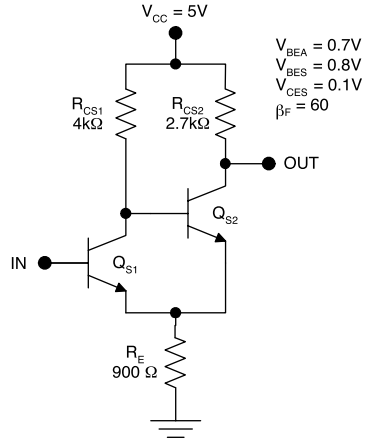


FIGURE 15.42
Emitter-coupled Schmitt trigger (P15.5).

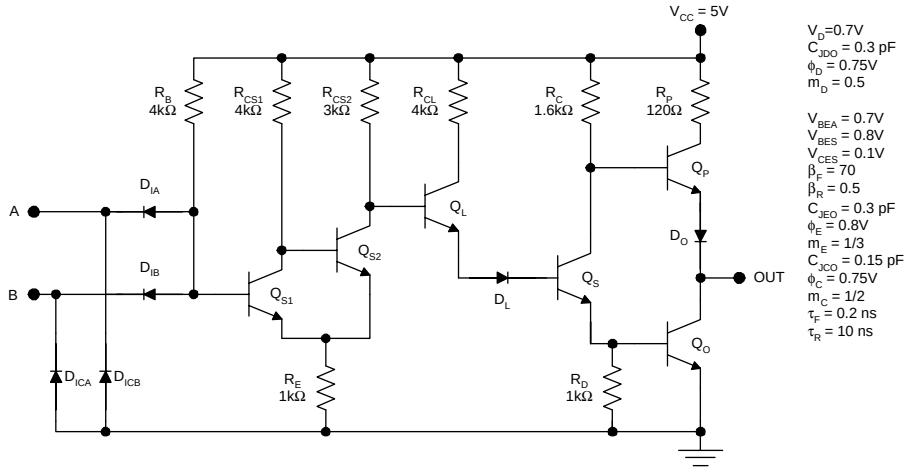


FIGURE 15.43
TTL Schmitt trigger (P15.6).

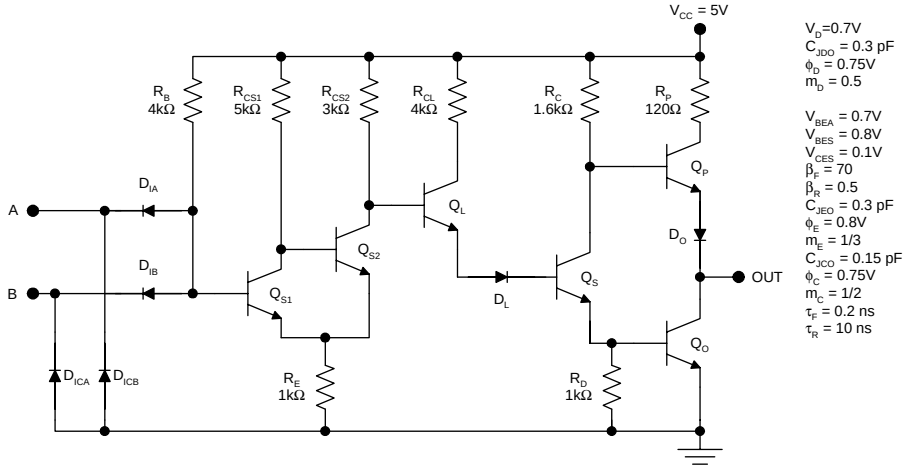


FIGURE 15.44
TTL Schmitt trigger (P15.7).

P15.6. Consider the TTL Schmitt trigger illustrated in Figure 15.43.

1. Estimate the trip voltages using hand calculations.
2. Determine the trip voltages using SPICE; compare the SPICE and hand-calculated values.

P15.7. Consider the TTL Schmitt trigger shown in Figure 15.44.

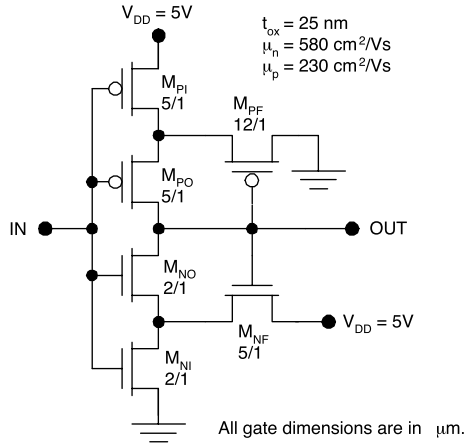


FIGURE 15.45
CMOS Schmitt trigger (P15.8).

1. Determine and plot the voltage transfer characteristic by hand calculations.
2. Repeat using SPICE. (Two DC sweeps are necessary.)
3. Plot the SPICE and hand-calculated results on one graph for comparison.

P15.8. Consider the CMOS Schmitt trigger illustrated in Figure 15.45. V_T = 0.6 V.

1. Estimate the trip voltages using hand calculations.
2. Determine the trip voltages using SPICE; compare the SPICE and hand-calculated values.

P15.9. Consider the CMOS Schmitt trigger shown in Figure 15.46. V_T = 0.6 V.

1. Determine and plot the voltage transfer characteristic by hand calculations.
2. Repeat using SPICE. (Two DC sweeps are necessary.)
3. Plot the SPICE and hand-calculated results on one graph for comparison.

P15.10. Design a standard TTL Schmitt trigger with 1.5-V hysteresis using a 5V supply. Provide a complete schematic with all resistor values and verify your design using SPICE.

P15.11. Design a low power Schottky TTL Schmitt trigger with 0.6 V hysteresis using a 5-V supply. The average DC dissipation must be less than 3 mW. Provide a complete schematic with all resistor values and verify the design using SPICE.

P15.12. Design a CMOS Schmitt trigger with 2.2-V hysteresis using a 5-V supply. t_{ox} = 13 nm, μ_n = 580 cm²/Vs, μ_p = 230 cm²/Vs, and L =

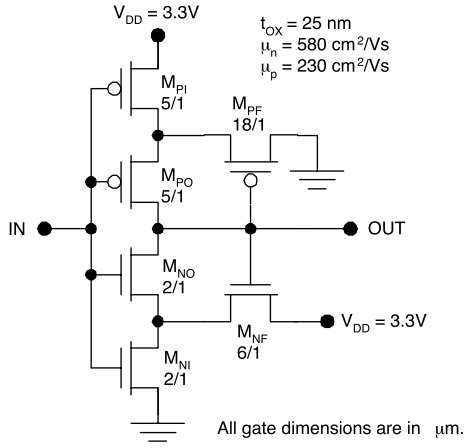


FIGURE 15.46
CMOS Schmitt trigger (P15.9).

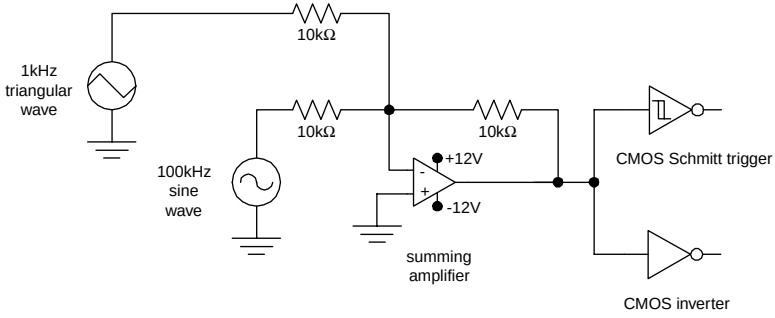


FIGURE 15.47
Circuit for testing Schmitt triggers (P15.13).

1 μm. Provide a complete schematic with all component values and verify the design using SPICE.

P15.13. The circuit of Figure 15.47 can be used to compare the noise rejection properties of two inverters. Consider the Schmitt trigger and inverter shown in Figure 15.48. Assume the CMOS inverter is symmetric with V_{DD} = 5 V and V_T = 0.6 V. Assume the CMOS Schmitt trigger is symmetric with V_H = 2 V.

1. Use SPICE to quantify the noise rejection capability of the Schmitt trigger. To do this use a 5-V supply for the Schmitt trigger. Set the peak-to-peak amplitude of the triangular wave to 5 V and the DC offset to 2.5 V. This simulates a slowly varying digital signal. Now increase the peak-to-peak amplitude of the sine wave source (which simulates noise) until bit

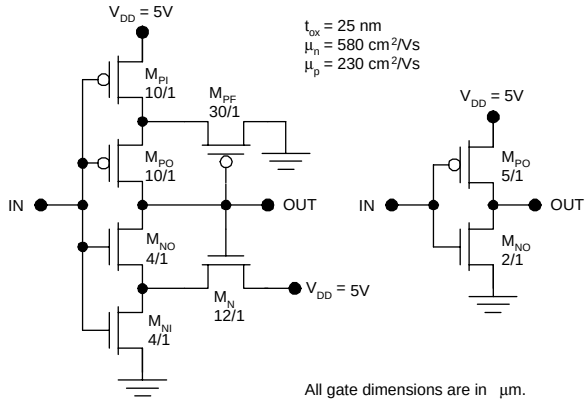


FIGURE 15.48
Schmitt trigger and inverter (P15.13).

errors are observed at the output. Determine this peak-to-peak value; is it what would be expected?

2. Use SPICE to quantify the noise rejection capability of the conventional inverter. To do this, use a 5-V supply for the conventional inverter. Set the peak-to-peak amplitude of the triangular wave to 5 V and the DC offset to 2.5 V. As in (a), increase the peak-to-peak amplitude of the sine wave source until bit errors are observed at the output. Determine this peak-to-peak value; is it what would be expected? How do the two circuits compare in their ability to reject noise?

References

1. Quad 3-state NOR R/S latches, Fairchild Semiconductor CD4043BC data sheet, www.fairchildsemi.com, 2002.
2. Octal D-type transparent latches and edge-triggered flip-flops, Texas Instruments SN54LS374 data sheet, www.ti.com, 2002.
3. AND-gated J-K master-slave flip-flops with preset and clear, Texas Instruments SN5472/SN7472 data sheet, www.ti.com, 1988.
4. Dual positive-edge-triggered D-type flip-flops, Texas Instruments SN54LV74A/SN74LV74A data sheet, www.ti.com, 2001.
5. CMOS Schmitt trigger — a uniquely versatile design component, Fairchild Semiconductor application note 140, www.fairchildsemi.com, 1975.
6. Quad 2-input NAND Schmitt trigger, Fairchild Semiconductor 74F132 data sheet, www.fairchildsemi.com, 2000.
7. www.cadence.com (Cadence).

16

Digital Memories

16.1 Introduction

Digital memories store bits of information for later use by processors, displays, and I/O devices. They are crucial elements in numerous products, including servers, personal computers, personal digital assistants (PDAs), digital cameras, and voicemail machines, to name a few.

Broadly, digital memories can be classified as volatile or nonvolatile. Nonvolatile memories retain their data when the system power is turned off; this kind of memory is used in digital cameras, PDAs, voicemail machines, and smart cards. Nonvolatile memory is also used to store the system software for microprocessor systems. Computers and servers need large amounts of volatile memory for temporary storage of data. This volatile memory is loaded with programs and files only while software is running. When not in use, these programs and files reside on mass media such as fixed or removable disks, which have greater capacity than the volatile memory but are considerably slower.

In large digital systems, data storage is organized according to capacity and speed. The highest capacity media (which are also the slowest) are farthest from the processor. These are nonvolatile storage media such as magnetic and optical disk drives. Next come nonvolatile memory circuits. Still closer to the processor are volatile memory circuits and, finally, additional memory circuitry is placed on the processor chip itself. This memory has limited capacity but is optimized for speed of access.

All digital memory circuits use the basic organization shown in Figure 16.1. The actual memory cells are arranged in a rectangular array with 2^N rows and 2^M columns. Such a scheme requires N + M address bits and provides 2^{N+M} cells, each of which may contain one or more bits. If each cell contains L bits, then the memory chip will have L data lines. The data in any individual cell may be accessed by selecting the j*th* row and the k*th* column. The row is selected by applying an N-bit row address, which is decoded by the row decoder. The column is selected by applying an M-bit address, which is decoded by the column decoder.

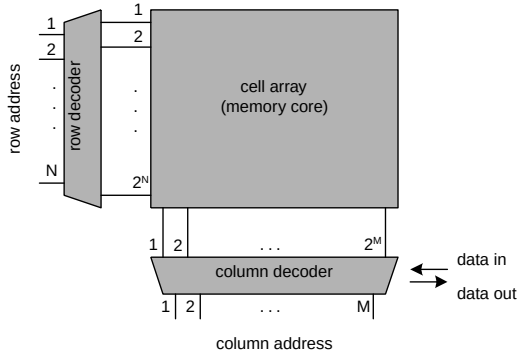


FIGURE 16.1
General layout of a digital memory.

The column decoding circuitry usually performs another function as well, that is, to transfer data in and out of the memory chip. Therefore, bits of data are transferred on column lines within the memory core. For this reason, the column lines are called bit lines while the row lines are called word lines. Often the core of the memory is split into several blocks to limit the lengths of the individual word and bit lines because the access time for a memory is often limited by the delays associated with the long interconnects serving as bit and word lines.

Memory chips are classified as read-only memory (ROM) or random access memory (RAM). The latter type should really be called "read–write memory" because it can write data, which distinguishes it from ROM. On the other hand, the name "RAM" is somewhat misleading because *both* ROM and RAM provide random access: the cells can be accessed in any random order.

Random access memories can be further classified as static RAM (SRAM) and dynamic RAM (DRAM). Because static RAMs store information in latches, these chips retain their data as long as the system power is on, without the need for clocking or refreshing. Dynamic RAMs store information using charges on capacitors. Because these capacitors exhibit some level of charge leakage, the voltages must be sensed and refreshed every few milliseconds to prevent data loss. RAMs are inherently volatile in nature.

Read-only memories can be further classified according to their capabilities for programming and erasing. Those circuits called simply "ROM" are factory programmed and may not be erased or reprogrammed after fabrication. Programmable read-only memory (PROM) may be programmed by the customer one time only; no provision is made for erasure or reprogramming. Erasable programmable read-only memory (EPROM) may be programmed, erased, and reprogrammed many times; however, erasure requires removing the chip from the system for flood exposure by ultraviolet radiation. Electrically erasable programmable read-only memory (EEPROM, or E^2PROM) is considerably more convenient because the erase and program operations may be done with the chip in place. "Flash memory" is a special type of

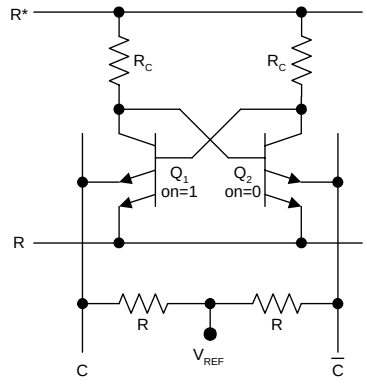


FIGURE 16.2
Emitter-coupled SRAM cell.

EEPROM that allows large blocks of data to be erased quickly. All ROMs are nonvolatile.

Today an almost limitless selection of memory circuits exists in the marketplace. No attempt will be made to catalog them here. Instead, the focus will be on the general principles underlying memory circuits.

16.2 Static Random Access Memory (SRAM)

Static RAMs use bistable latch circuits to store bits of data. These latches may be built using bipolar transistors or MOSFETs and numerous designs exist.[1–8]

Emitter-coupled SRAM cells are latches constructed from cross-coupled BJT inverters. One such circuit is shown in Figure 16.2. The circuit takes its name from the fact that data are transferred to and from column lines using the extra emitters in the bipolar transistors. The design shown uses two row lines and two column lines per cell. Under standby conditions, the latch will remain in one of its two stable states indefinitely as long as the power is not interrupted. Q_1 "on" represents the storage of a "one" and Q_2 "on" represents the storage of a "zero."

Under standby operation, the voltage difference between the R* and R lines is made just sufficient to maintain the state of the latch (about 1 V). V_{REF} is chosen so that, under standby conditions, the emitter–base junctions coupling to the column lines are reverse biased.

For a read operation, the R* line is brought high ($\sim V_{CC}$) while the R line is brought high ($\sim V_{CC}/2$). This increases the current flowing in the "on" transistor. At the same time, this transistor drives its column line because the voltage on the R line is greater than V_{REF}. Then, the voltage difference between

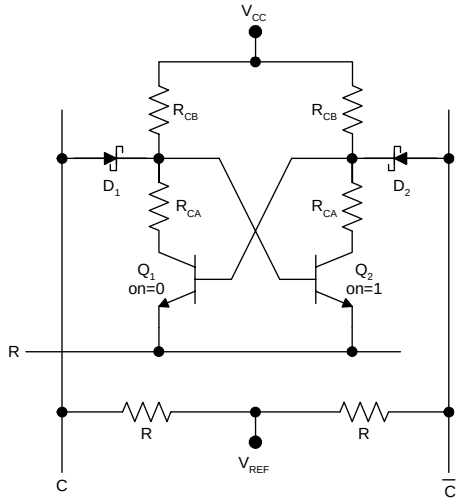


FIGURE 16.3
Diode-coupled SRAM cell.

the column lines is sensed; greater voltage on C indicates that Q_1 is conducting and the stored bit is a one. If the voltage on $\overline{C}$ is greater, the stored bit is a zero. For a "write" operation, the R* line is brought high ($\sim V_{CC}$) while the R line is brought high ($\sim V_{CC}/2$). To write a one, the C line is brought low, forcing Q_1 to turn on. To write a zero, $\overline{C}$ is brought low, forcing Q_2 to turn on.

Another SRAM cell based on cross-coupled BJT inverters is shown in Figure 16.3. This diode-coupled SRAM circuit takes its name from the fact that data are transferred to and from column lines using Schottky diodes. Q_1 "on" represents the storage of a zero and Q_2 "on" represents the storage of a one.

Under standby operation, the voltage difference between V_{CC} and the R line is made just sufficient to maintain the state of the latch (about 1 V). V_{REF} is chosen so that, under standby conditions, the diodes coupling to the column lines are reverse biased.

For a read operation, the R line is brought low (~0) and the voltage difference between the column lines is sensed. Lesser voltage on C indicates that Q_1 is conducting and the stored bit is a zero; if the voltage on $\overline{C}$ is less, the stored bit is a one. For a write operation, the R line is brought low (~0) and the column lines are driven appropriately. To write a one, the C line is brought high, forcing Q_2 to turn on. To write a zero, $\overline{C}$ is brought high, forcing Q_1 to turn on.

SRAMs based on bipolar transistors have low density and high dissipation. The low density results from the use of bipolar transistors, which are larger than MOSFETs, and the use of resistors, which are area intensive. The dissipation is relatively great per cell despite efforts to reduce the row voltages

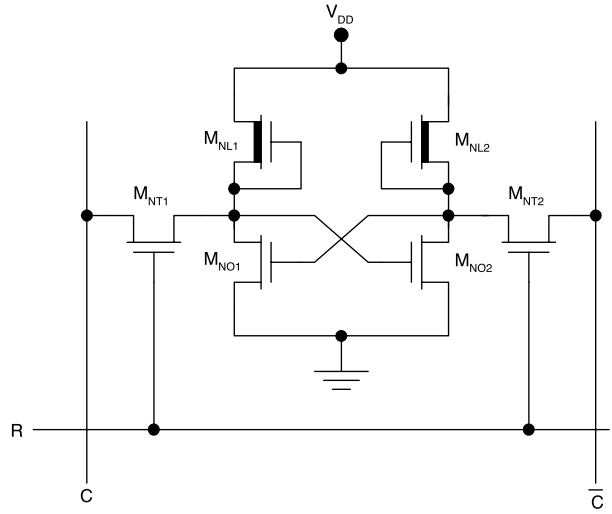


FIGURE 16.4
NMOS 6T SRAM cell.

under standby conditions. This stems from the fact that bipolar transistors are current-controlled rather than voltage-controlled devices. Denser SRAMs with reduced dissipation may be realized using NMOS, CMOS, or BiCMOS.

An NMOS six-transistor (6T) SRAM cell is shown in Figure 16.4. This cell comprises two cross-coupled NMOS inverters and two access transistors, M_{NT1} and M_{NT2}. Under standby conditions, the R line is kept low (~0), so the access transistors are cut off. For a read operation, the R line is brought to V_{DD}, turning on the access transistors, and the difference between the column lines is sensed. Greater voltage on C indicates that M_{NO2} is conducting and the stored bit is a one; if the voltage on $\overline{C}$ is greater, the stored bit is a zero. For a write operation, the R line is brought to V_{DD}, turning on the access transistors, and the column lines are driven appropriately. To write a one, C is driven to V_{DD}; to write a zero, $\overline{C}$ is driven to V_{DD}. The six-transistor (6T) CMOS SRAM cell is very similar, as shown in Figure 16.5. It comprises two cross-coupled inverters and the read–write operations are the same as for the NMOS 6T SRAM cell.

The sense amplifiers used in CMOS SRAMs vary in design but they are all essentially NMOS, CMOS, or BiCMOS differential amplifiers. An example of a CMOS sense amplifier is shown in Figure 16.6. When the clock signal is low, M_{NCLK} is cut off and the nodes at the drains of the differential pair transistors (M_{ND1} and M_{ND2}) precharge to V_{DD}. A read operation is initiated when the clock goes high, turning on M_{NCLK}. Then, if the voltage on the column line C is greater than that on $\overline{C}$, M_{ND1} will turn on, M_{ND2} will turn off, and the output will go high. If instead the voltage on C is less than that on $\overline{C}$, M_{ND1} will turn off, M_{ND2} will turn on, and the output will go low. The

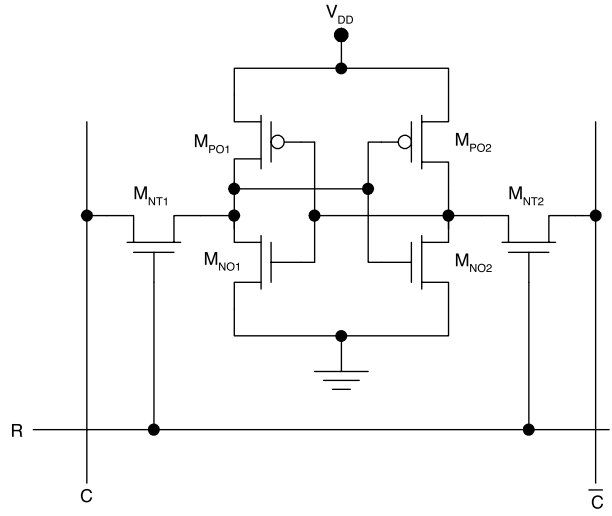


FIGURE 16.5
CMOS 6T SRAM cell.

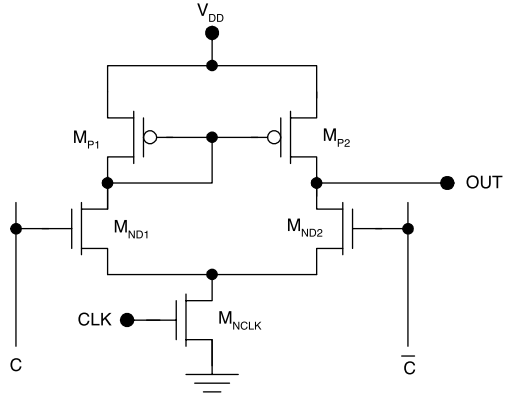


FIGURE 16.6
CMOS SRAM sense amplifier.

read time is determined in great part by the voltage gain of the differential amplifier and the rise and fall times of the signals on the bit lines C and $\overline{C}$. Thus two or more stages of differential amplifiers may be used to improve the overall gain and, consequently, the read time. The availability of complementary outputs at the drains of M_{ND1} and M_{ND2} facilitates cascading in this way.

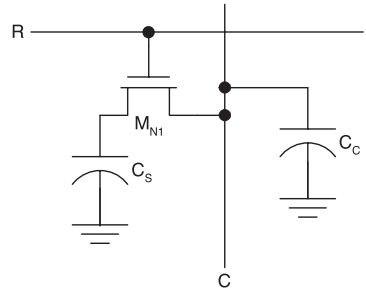


FIGURE 16.7
DRAM cell involving one transistor and one storage capacitor (1T1C cell), connected to a column capacitance.

16.3 Dynamic Random Access Memory (DRAM)

MOS SRAM requires a minimum of six transistors per bit. Dynamic random access memories (DRAMs)[1–4,9–29] achieve much higher density because they are constructed with three or fewer transistors per bit. The so-called "1T1C" DRAM cell comprises one n-channel MOSFET and a storage capacitor (Figure 16.7). The right-hand capacitor in the figure is the column capacitance and is not associated with any one cell. The bit stored in this cell is represented by the voltage on the storage capacitor C_S. If the voltage on the storage capacitor is V_{DD}, a one is stored; logic zero is represented by 0 V. At the present time, dedicated DRAM chips use 1T1C cells exclusively because of their high density. On the other hand, DRAMs embedded in systems on a chip (SOCs) often employ two-transistor cells. These cells are less dense, but provide improved signal-to-noise ratios and therefore greater reliability in the presence of cross talk.

The operation of the 1T1C cell of Figure 16.7 is as follows. During a write operation, the row line is brought high to turn on the access transistor and the column line is forced to V_{DD} (for logic one) or zero (for logic zero). During a read operation, the access transistor is turned on and the column voltage is sensed. If the column voltage rises above $V_{DD}/2$, then a one is inferred. If the column voltage drops below $V_{DD}/2$, then a zero is read. Typically, the column capacitance is much larger than the cell storage capacitance so that only small voltage excursions are obtained on the column line. High-gain amplifiers (called sense amplifiers and contained in the column circuitry) are used to read these weak signals. Once the sense amplifier makes a decision regarding the value being read, it forces the column line to V_{DD} or zero to refresh the signal on the storage capacitor. After this signal is refreshed, the access transistor may be turned off and the read operation is

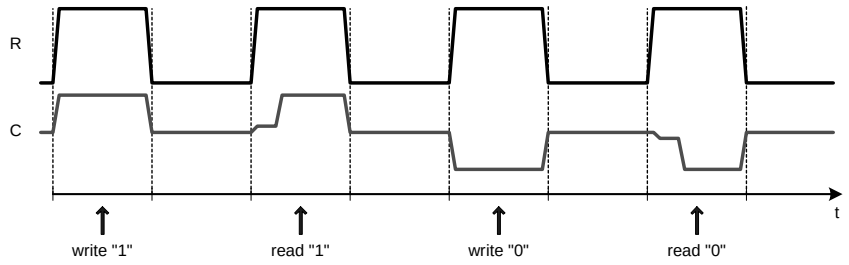


FIGURE 16.8
DRAM read and write operations.

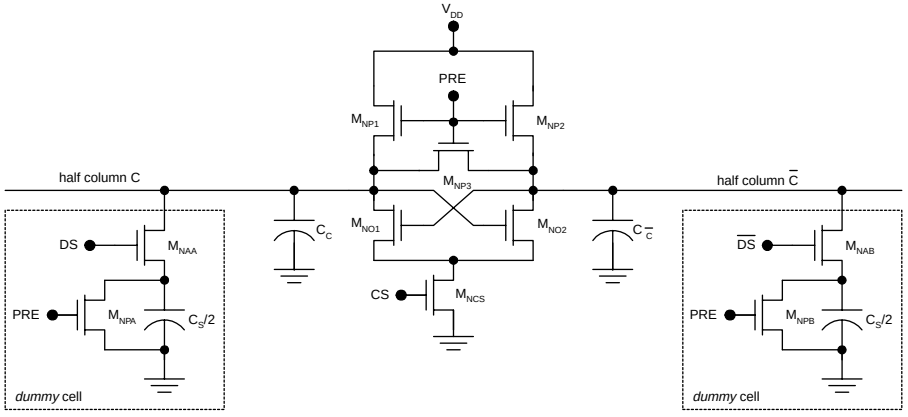


FIGURE 16.9
DRAM read/refresh circuit.

complete. The read and write operations for the one-transistor DRAM cell are illustrated in Figure 16.8.

The data stored in 1T1C DRAM cells must be refreshed periodically because of small leakage currents in the storage capacitors and access transistors. Typically, the leakage is of such magnitude that the data must be refreshed every few milliseconds. Fortunately, this function is built into the memory chip and is transparent to the end user.

There are numerous designs for 1T1C DRAM read–refresh circuits; one such circuit is shown in Figure 16.9. This DRAM read–refresh circuit utilizes an NMOS latch circuit with two reference cells (*dummy cells*). The column is split into two half columns, C and $\overline{C}$. Half column C is connected to half of the cells in the column and half column $\overline{C}$ is connected to the other half of the cells in the column. Each half column has an associated half-column capacitance (C_C and $C_{\overline{C}}$).

The read–refresh procedure is as follows. First, PRE goes high, precharging the half-column capacitances to V_{DD} and discharging the dummy cell capacitances to zero. Next, the row is selected (for the cell to be read and refreshed). The decoding circuitry is designed such that the dummy cell on the opposite

side of the circuit is selected by bringing *DS* or $\overline{DS}$ high. (For example, if the row selected is in half column C, then $\overline{DS}$ is brought high.) Finally, the column is selected by bringing *CS* high and a small voltage difference between the two half columns will be amplified by the high loop gain of the latch. If a cell in half column *C* that stored *logic one* is read, then the voltage on half column *C* will be *greater* than the dummy cell voltage on half column $\overline{C}$; the latch will force *C* to V_{DD} and $\overline{C}$ to GND. If a cell that stored *logic zero* in half column *C* is read, then the voltage on half column *C* will be *less* than the dummy cell voltage on half column $\overline{C}$; the latch will force *C* to GND and $\overline{C}$ to V_{DD}.

There are many different designs for 1T1C DRAM cells. Their differences relate to the physical design of the storage capacitors and access transistors. Numerous fascinating schemes have been devised to shrink the footprint of the capacitor and therefore increase the density of the resulting DRAM. Many of these involve folding the capacitor in vertical structures such as trenches; some involve using multiple arms or cylinders. Remarkably, it has proven possible to fabricate such structures with excellent levels of circuit yield.

Today DRAMs are the digital circuits with the highest level of integration. This is a consequence of two important factors. First, the core of a DRAM contains many simple and identical cells. It is therefore common to design in a high degree of redundancy, which minimizes the impact of defects on the circuit yield. Other types of circuits, such as microprocessors, do not enjoy this advantage. Second, customer demands for computer memory have fueled the drive for denser DRAMs. These market forces are so strong that the DRAM business (along with the microprocessor- and application-specific IC industries) has been a technology driver for the entire silicon integrated circuit industry. Thus the requirements for DRAMs set the pace of development for lithographic tools and fabrication equipment.

Relentless device scaling will continue to increase the densities of DRAM chips for some time to come. Further increases will also come about with steady increases in die sizes, made possible by the reduction in process defects. Eliminating the capacitor, resulting in a true 1T cell, has also aroused considerable interest. Such a cell requires the storage of charge in the MOSFET structure and imposes considerable challenges; nevertheless, the 1T cell offers potential for further increases in the capacity of DRAM chips.

16.4 Read-Only Memory (ROM)

Read-only memory (ROM) is programmed at the factory and cannot be written to or reprogrammed once installed. Circuits of this type are entirely customized to the needs for a single product made by a single manufacturer. ROMs are usually fabricated with a single diode or transistor per bit, so high density and low power are achieved.

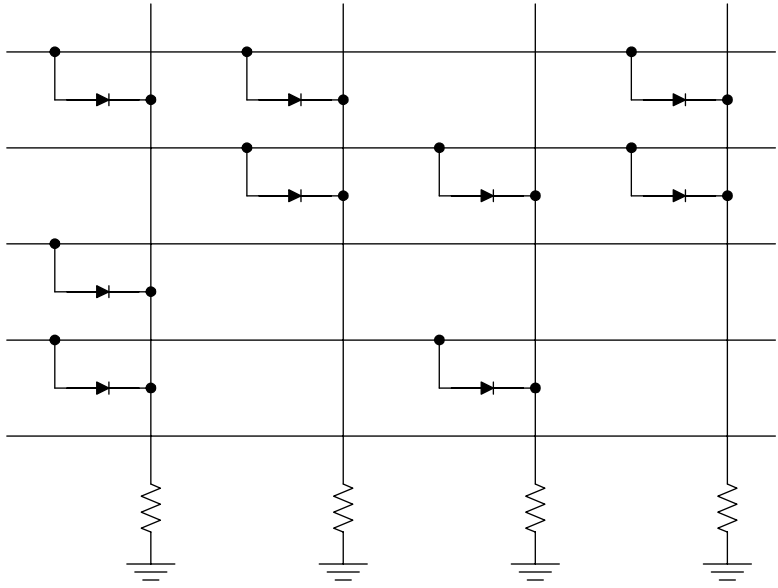


FIGURE 16.10
Diode ROM.

In a diode ROM such as the one shown in Figure 16.10, logic one is programmed by placing a diode at the junction of the row and column lines and logic zero is programmed by the absence of a diode. During the read operation, the row line is brought high; the presence of a diode causes the associated column line to go high. On the other hand, the absence of a diode allows the associated column line to go low. A pull-down resistor between the column line and ground ensures that the column line is pulled to ground in this case.

In practice, diode ROMs are fabricated with a diode at each row–column junction, which is then wired to the row and column lines only as needed. This approach has two advantages. First, the only custom step is the final metallization that wires the diodes; therefore, large batches of wafers can be processed with all steps up to the (custom) final metallization. The economy of scale reduces the total cost. Second, it is possible to store large quantities of the partially fabricated wafers, thus reducing the order-to-delivery lead time for the customer.

In a BJT ROM, logic one is programmed by placing a bipolar transistor at the junction of the row and column lines as shown in Figure 16.11. Logic zero is programmed by the absence of a transistor. During the read operation, the row line is brought high. The presence of a BJT emitter follower causes the associated column line to go high and the absence of a transistor allows the associated column line to go low. As with diode ROM, the programming

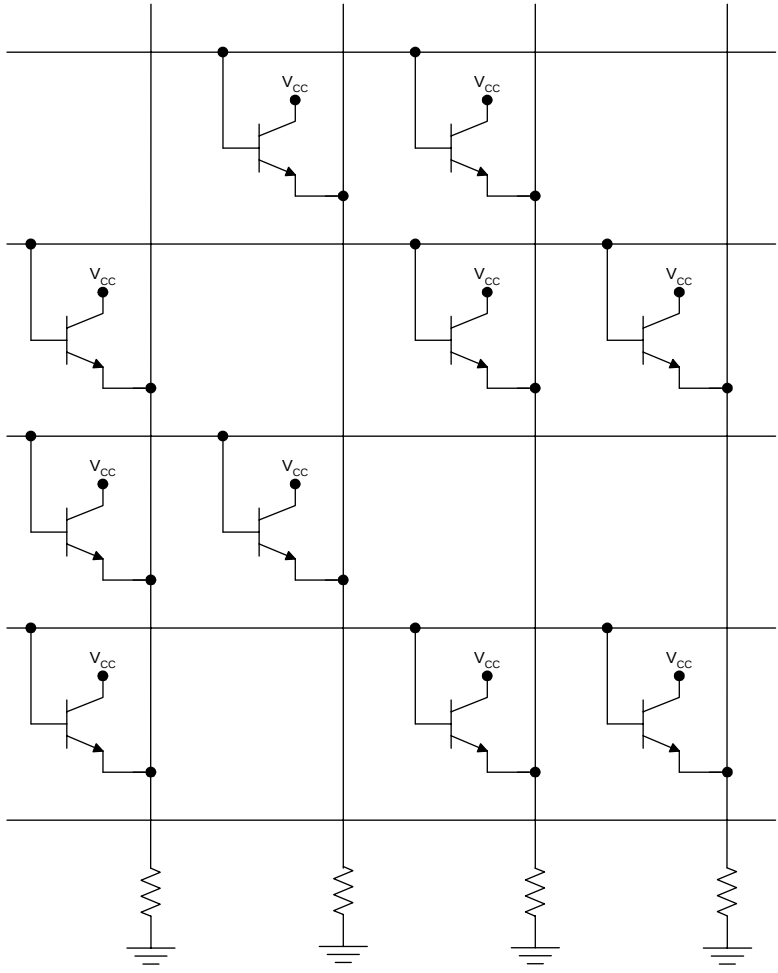


FIGURE 16.11
BJT read-only memory.

is done in the final metallization step, with the same benefits. An advantage of the BJT design is that the emitter followers provide lower output impedance and better dynamic performance than for the case of the diode ROM. A disadvantage is the need to route V_{CC} lines to all of the transistors.

NMOS ROM can be realized using the design shown in Figure 16.12, in which each column line has a depletion type n-channel pull-up transistor. Logic one is programmed by placing an enhancement type MOSFET between the row and column lines. Because each column forms an NMOS NOR gate, this design is referred to as NMOS NOR ROM. As with the bipolar ROMs, it is usual practice to carry out the programming in the final metallization

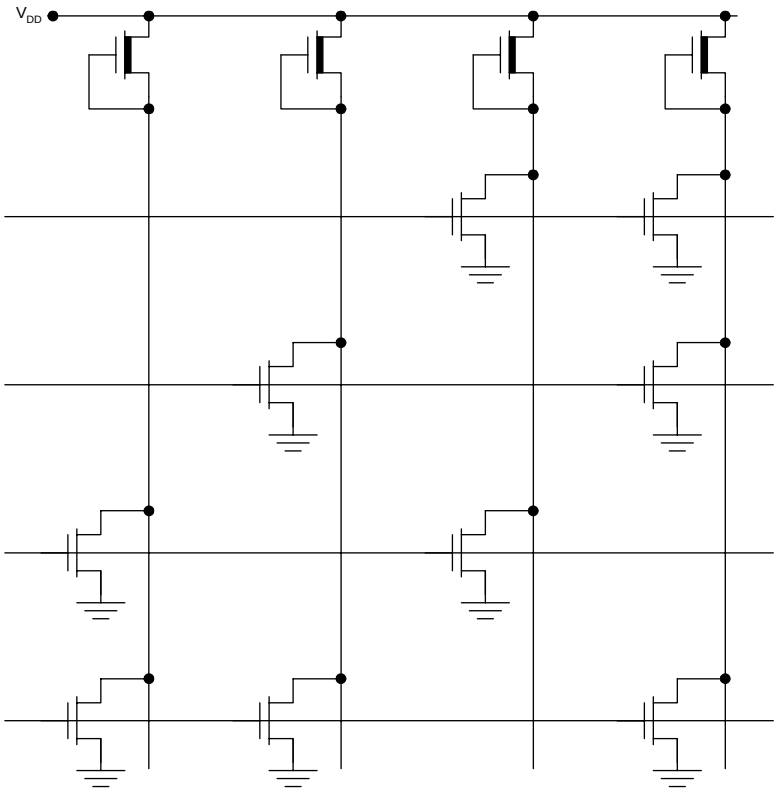


FIGURE 16.12
NMOS NOR read-only memory.

step to minimize cost and lead time. In contrast to bipolar ROMs, NMOS ROM is denser because of the reduced area of MOSFETs compared to diodes or bipolar transistors.

CMOS NOR ROM can also be realized with high density and exhibits reduced power dissipation compared to the NOR design. In a CMOS NOR ROM, the pull-up devices are p-channel MOSFETs, as shown in Figure 16.13. These p-channel MOSFETs are clocked by a precharge signal (labeled V_{PRE} in the illustration). Prior to a read operation, the precharge signal is momentarily brought low to turn on all of the p-channel MOSFETs. This serves to charge the column capacitances to V_{DD}. Following this, a row line is brought high, thus turning on any n-MOSFETs associated with that row, which will in turn pull down the associated columns.

Notice that the CMOS ROM comprises an array of dynamic CMOS NOR gates, one on each column. This leads to the name CMOS NOR ROM. Also, the dynamic operation leads to greatly reduced dissipation compared to the NMOS NOR ROM.

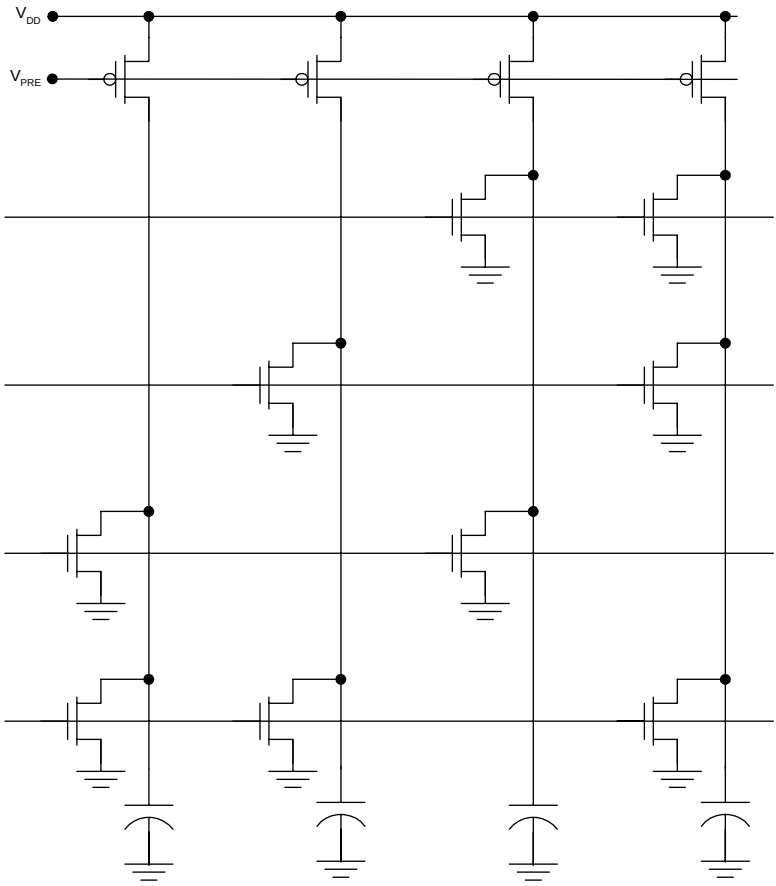


FIGURE 16.13
CMOS NOR read-only memory.

16.5 Programmable Read-Only Memory (PROM)

The requirement of factory programming renders ROM unpractical for many applications. For small-volume products, factory customized ROMs are simply uneconomical and, for prototypes, they are too slow. These applications indicate programmable read-only memories (PROMs)[3] that are programmed *after* manufacture.

One way of achieving this programmability is with microfuses, as shown in the bipolar PROM of Figure 16.14. The architecture is essentially the same as in the BJT ROM, but a transistor is provided for each cell. Logic zero is

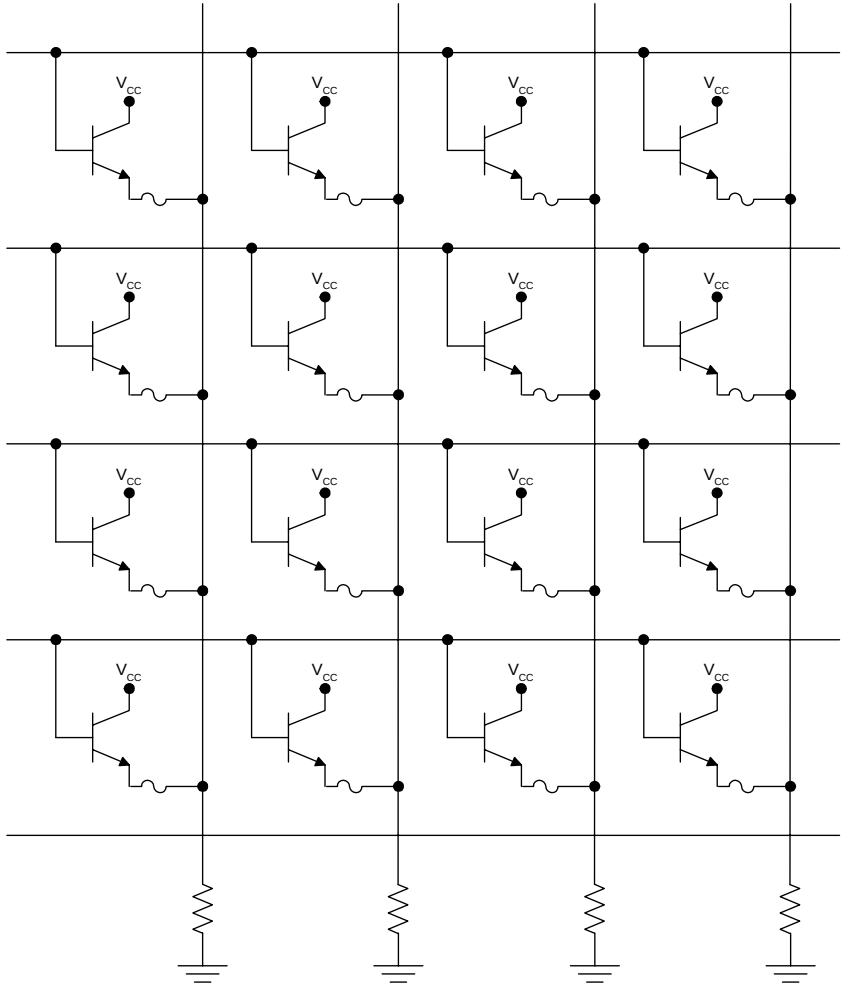


FIGURE 16.14
BJT programmable read-only memory.

programmed (irreversibly) by blowing the fuse associated with a particular transistor. Microfuses are typically made using a metal such as nichrome. During programming, transistors are turned on systematically with sufficient electrical current to vaporize the microfuse, resulting in an open circuit. During normal operation the supply voltage is kept low enough to prevent unintentional programming.

Because this design approach may be applied to NMOS and CMOS PROMs, these MOS ROMS also use microfuses.

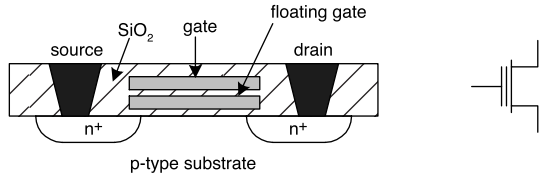


FIGURE 16.15
Floating gate avalanche injection MOS transistor (FAMOS).

16.6 Erasable Programmable Read-Only Memory (EPROM)

A disadvantage of PROM is that it can be programmed only once; changing a single bit requires a fresh circuit — a serious drawback for prototyping and product development. Fortunately, erasable programmable read-only memory (EPROM) chips allow multiple erase–program cycles.

The basis for the EPROM is a specially designed MOSFET called a floating gate avalanche injection MOS transistor, or FAMOS, which is shown in Figure 16.15. As the name suggests, this device has a second gate that floats (is not electrically connected to the rest of the circuit). The placement of electrical charge on the floating gate shifts the threshold voltage of the transistor. Negative charge placed on the floating gate repels electrons in the channel, making the threshold voltage more positive. Positive charge on the floating gate attracts electrons to the channel and makes the threshold voltage more negative. This property of the FAMOS is exploited in PROMs.

The circuit design for an EPROM is shown in Figure 16.16. This is essentially an NMOS NOR ROM constructed using floating-gate MOSFETs. Logic zero is programmed by making the threshold voltage of the associated FAMOS greater than V_{DD}, ensuring that the transistor will never turn on. A FAMOS programmed in this way behaves as if it has been removed from the circuit. Logic one is programmed by removing all charge from the floating gate. The devices are designed so that this results in a threshold voltage greater than 0 and less than V_{DD}.

Blanket erasure of the EPROM is achieved by flooding the integrated circuit with ultraviolet radiation while grounding the sources of the FAMOS devices. The ultraviolet radiation renders the gate oxide slightly conductive, allowing all electrical charge to leak away from the floating gates. The FAMOS devices are designed so that their threshold voltages are positive, but much less than V_{DD}, with zero charge on the floating gate. Therefore, this blanket erasure resets each bit to logic zero. The erasure operation is done by shining the ultraviolet radiation through a special plastic window in the top of the chip package, giving the EPROM an unmistakable appearance. Typically, this process takes several minutes to complete.

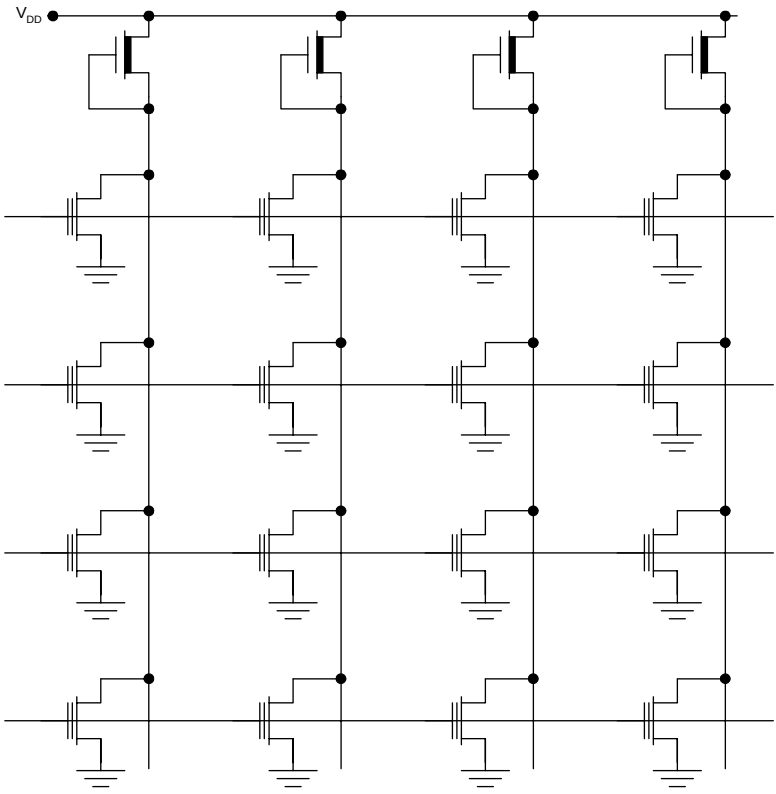


FIGURE 16.16
Erasable programmable read-only memory (EPROM).

Programming is done by the selective adjustment of the threshold voltages for the FAMOS devices. Negative charge is placed on the floating gate of each FAMOS where logic one must be stored. This is achieved by grounding the column line and placing large positive voltages on the row line and the V_{DD} supply line. This causes the selected FAMOS to operate in the avalanche breakdown mode. In this mode of operation, electrons drifting from the source to the drain become sufficiently energetic* to be injected through the gate insulator to the floating gate. This is a self-limiting process, so the amount of negative charge placed on the floating gate is determined by the voltages placed on the column and V_{DD} lines. The process is designed so that the end result is a threshold voltage greater than V_{DD}. This effectively disables the associated FAMOS to result in logic one at the selected location.

* These high-energy electrons are called "hot electrons" because they are not in thermal equilibrium with the semiconductor lattice. Their energy corresponds to the equilibrium energy of electrons in a lattice at temperatures of 1000 to 10,000 K.

Two drawbacks are associated with EPROMs: 1) the inconvenience of erasure, which requires removal of the chip from the system to an ultraviolet flood chamber and 2) the inevitable oxide degradation that occurs with repeated erase–program cycles. Induced by the injection of hot electrons, this degradation gradually renders the oxide leaky and limits the total number of erase–program cycles to hundreds.

16.7 Electrically Erasable Programmable Read-Only Memory (EEPROM)

The inconvenience of erasing EPROMs has led to the development of electrically erasable programmable read-only memories (EEPROM, or E^{2}PROM).[30] The EEPROM is based on the FLOTOX transistor illustrated in Figure 16.17. This is a specially fabricated floating gate MOSFET that can be erased and programmed electrically by quantum mechanical tunneling.

Quantum mechanical tunneling allows electrons to cross a thin barrier such as an insulator. If the electron wave function takes on a finite value on the opposite side of the barrier, a finite probability exists that the electron will spontaneously appear there. However, the wave function of an electron decays rapidly with distance, so tunneling is only possible with barriers less than about 10 nm thick. The FLOTOX transistor is specially designed to have a region of thin tunneling oxide over the drain that allows transport of electrons to and from the floating gate. During erasing or programming, electrons tunnel through this oxide in a number of intermediate jumps involving defect states in the oxide. The specific process involved is called Fowler–Nordheim tunneling.

The EEPROM circuit requires two transistors per bit as shown in Figure 16.18. Each cell contains one conventional MOSFET as well as a FLOTOX device. To program logic one, sufficient negative charge is placed on the floating gate so that the threshold voltage ends up greater than V_{DD}, thus ensuring that the FLOTOX device will never turn on. Then, during a read operation, bringing the row line high will turn on the access transistor (the conventional MOSFET), but the voltage on the column line will remain at V_{DD} if the FLOTOX device is cut off. In practice, logic one is programmed by placing a positive voltage on the row (gate) and a negative voltage on the column (drain), causing electrons to tunnel from the drain to the floating gate.

To program logic zero, the opposite biasing is used. A positive voltage is applied to both the row (gate) and the column (drain), causing electrons to tunnel from the floating gate to the drain. This renders the threshold voltage of the FLOTOX device less than V_{DD}. When thus programmed, the FLOTOX device operates in the linear mode with V_{DD} applied at its gate. Therefore, when the row line is brought high for a read operation, the access transistor and the FLOTOX device are linear, bringing the column line to ground.

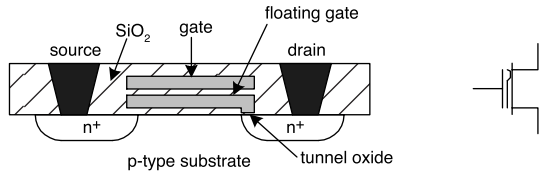


FIGURE 16.17
FLOTOX transistor and circuit symbol.

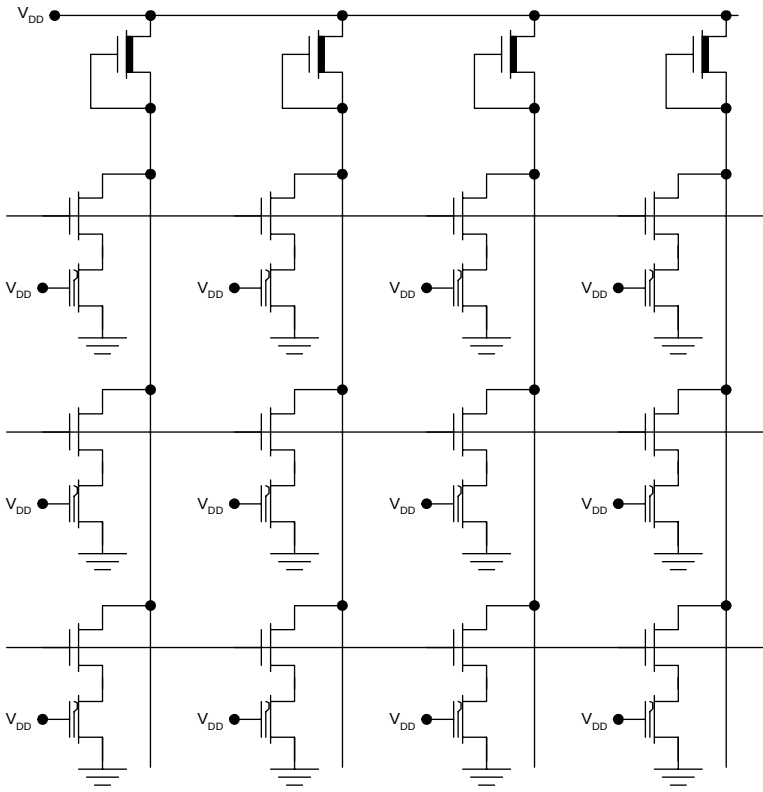


FIGURE 16.18
Electrically erasable programmable read-only memory (EEPROM).

The EEPROM needs two transistors per bit because it is not possible to control the programmed threshold voltage of the FLOTOX transistor tightly. After programming logic zero, the threshold voltage of the FLOTOX transistor may become *negative*, transforming it into a *depletion type* device. As a result, placing the FLOTOX device on the node by itself would pull the column line down to ground *even when the row was not selected*.

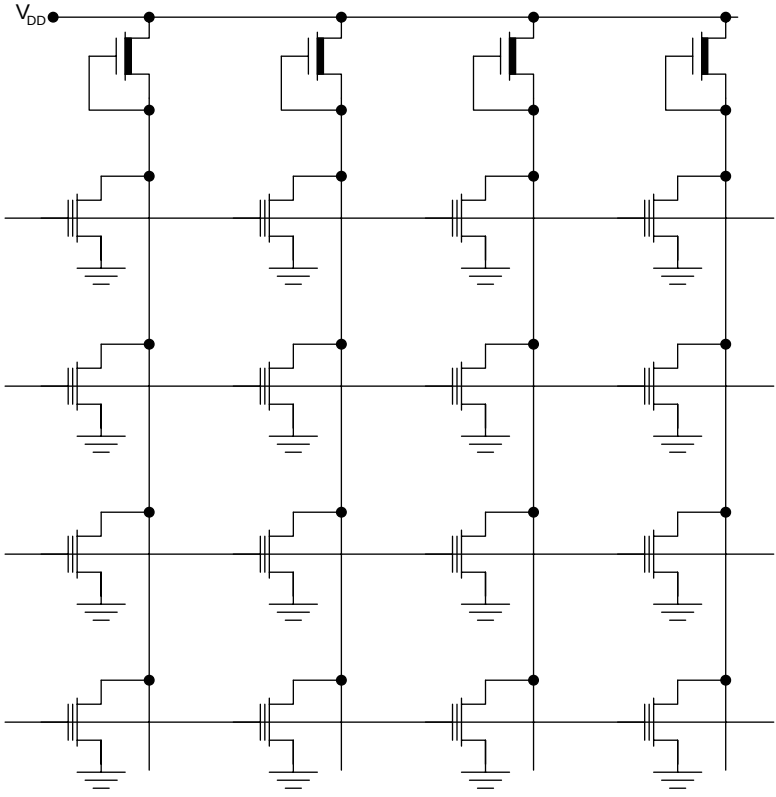


FIGURE 16.19
Flash memory.

16.8 Flash Memory

Flash memory[1–4,31–40] combines the flexibility of EEPROM with the high density of EPROM. This type of memory uses ETOX devices, which are similar to FAMOS transistors except that they have a thin tunneling oxide under the floating gate. The flash memory circuit is illustrated in Figure 16.19.

The programming is done by the avalanche injection of electrons, similar to the case for FAMOS transistors. Erasure is accomplished by Fowler–Nordheim tunneling, similar to the FLOTOX transistors. However, erasing is done in mass (by large blocks of memory), which allows end-of-process monitoring to ensure that the threshold voltages do not end up negative. Thus, flash memory can use a single transistor per bit, yielding roughly twice the density of EEPROM. Flash memory is used extensively in "smart cards" and personal multimedia products.

Recent innovations in flash memory include the replacement of the polysilicon floating gate with silicon nitride (Si_3N_4), and the incorporation of dual gates. Both approaches have made it possible to store two bits per cell in commercial flash memory chips at the present time. In principle, it should be possible to store more than two bits per cell, thereby increasing overall chip capacity.

16.9 Access Times in Digital Memories

Usually the access time in a digital memory is limited by the delay times for the interconnects forming the row and column lines. This is because these delays can greatly exceed the propagation delays for the circuits in the row and column decoders.

Consider a read operation. After the output of the row driver rises, the signal must propagate along the row line to the column to be selected. In the worst case, the column at the other end is to be selected, so the delay associated with the entire length of the row line must be considered. Once the row line has settled, the selected memory cell will swing the column line voltage in the positive or negative direction. Then, even if the rise or fall time is negligible at the selected cell, the delay associated with the column interconnect must be considered. In the worst case, the selected cell is at the opposite end of the column from the decoder circuitry, so the entire length of the column line must be considered. In MOS cells, an additional delay is associated with the cell driving the column capacitance. Bipolar memories are less susceptible to column loading because of their inherently better current driving capability. (As a rule of thumb, bipolar transistors provide four times greater current drive than MOSFETs for the same transistor area). However, the interconnect delay is usually still dominant, so it is adequate to consider only the interconnect delays in first-order calculations.

Suppose that a digital memory is organized with 2^N rows and 2^M columns. Suppose the row (word) line has a resistance per cell of R_w and a capacitance per cell of C_w. Suppose the column (bit) line parasitics are R_b and C_b. Then the access time can be estimated by summing the worst-case Elmore delays for the word and bit lines:

$$\begin{aligned} t_{read} &= t_{word} + t_{bit} \\ &= \ln(2) R_w C_w \frac{2^N + 1}{2} + \ln(2) R_b C_b \frac{2^M + 1}{2}. \\ &\approx \ln(2) R_w C_w 2^{N-1} + \ln(2) R_b C_b 2^{M-1} \end{aligned} \tag{16.1}$$

This analysis assumes that no repeaters have been used. However, repeaters are often used in the row lines of memories to reduce t_{word}.

The write time also involves the interconnect delays and can sometimes be estimated in the same fashion as the read time. More often, however, the write time involves other important contributions related to the physics of erasing or storing data in the cell. In flash memory, for example, the write operation typically takes 50 times as long as the read operation. In some cases, the process of reading destroys the data. Then every read operation must be followed by a write operation, which will increase the read time significantly. Such is the case for ferroelectric random access memories.

16.10 Emerging Memory Concepts

Flash memories are well suited to applications such as digital cameras, MP3 players, and wireless phones in which nonvolatile memory is a must and the write frequency is on the order of once per day. However, the maximum number of erase–write cycles for flash memory at the present time is only 10^6. This precludes using flash memory in digital computers because it would exceed this number of cycles in a short time. Instead, dynamic random access memory (DRAM) is used; however, the drawback of DRAM is that the data are lost upon power-down. Nonvolatile memory that can withstand unlimited erase–write cycles is needed. Recently, new types of memories have emerged that satisfy these requirements: *ferroelectric random access memory (FRAM)*,[41–49] *magnetoresistive random access memory (MRAM)*,[1,50–52] and *ovonic unified memory (OUM)*.[53–56] These technologies represent a major departure from today's digital memory devices because *they utilize the properties of nonsilicon materials.*

In FRAMs, each bit is stored in a ferroelectric capacitor connected to an access transistor, as shown in Figure 16.20. The capacitor is made using a ferroelectric material such as lead zirconium titanate (PZT). In the ferroelectric material, a built-in electric field is determined by the polarity of electric domains in it. Because these domains can be flipped in polarity by applying the appropriate bias voltage, the two possible polarities of the domains can be used to represent logic one and logic zero.

To store a bit in the FRAM cell of Figure 16.20, the word line, *R*, is brought high to turn on the access transistor and the bit line, *C*, is brought low or high to store a zero or one, respectively. The drive line *R'* is driven in complementary fashion to the bit line to provide the required bias polarity for the capacitor. This orients the domains in the ferroelectric capacitor so that a negative or positive voltage is stored in this capacitor to represent zero or one, respectively.

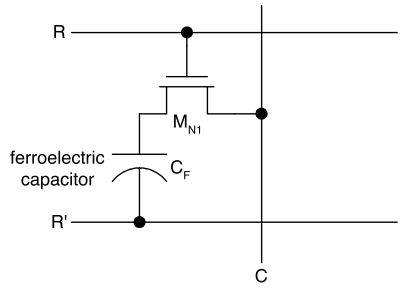


FIGURE 16.20
FRAM cell involving one transistor and one ferroelectric capacitor (1T1C cell).

To read a bit from the FRAM cell, the word line, *R*, and the column line, *C*, are brought high. A positive voltage pulse is applied to the drive line, *R'*, and the size of the resulting current pulse is used to determine the initial polarity of the voltage across the ferroelectric capacitor. Reading the bit erases the data in the cell, so refreshing is necessary after each read operation. In principle, FRAM cells should be able to endure up to 10^{16} erase–write cycles. At the present time FRAM products are capable of $>10^{12}$ cycles, a million-fold advantage over flash memory. Because the FRAM uses a cell similar to that for the DRAM, very high densities should be possible as the technology matures. The compatibility of the fabrication with conventional DRAM or CMOS processes is also an important advantage.

In the MRAM, each bit is stored in a small magnet made of a ferromagnetic material composed of magnetic domains that tend to line up with an externally applied magnetic field. Therefore, the direction in which the magnetic domains are aligned may be used to store a one or a zero. Readout of the bit can be accomplished using a tunnel junction.

Each cell of the MRAM contains a single access transistor and a magnetic tunneling junction (MTJ). Figure 16.21 shows such a magnetic tunneling junction with its circuit symbol. The tunneling junction comprises a thin (~2 nm) insulator such as aluminum oxide sandwiched between two layers of ferromagnetic material. In one of the ferromagnetic layers, the alignment of the ferromagnetic domains is fixed. In practice, this can be achieved using an antiferromagnetic (AF) pinning layer such as FeMn or IrMn with an intermediate layer of Ru. This combination creates a synthetic antiferromagnet (SAF). In the top ferromagnetic layer (the *free layer*), the domain alignment can be flipped by the application of simultaneous currents in the bit line and the digit line. The resistance of the MTJ is low if the two ferromagnetic layers have parallel domains but much higher if the domains are antiparallel.

The basic MRAM cell is shown in Figure 16.22; two row lines are necessary to allow programming. The simultaneous application of currents in the bit line and the digit line fixes the magnetic field of the free layer in the MTJ parallel to that in the fixed layer. Application of the currents with opposite

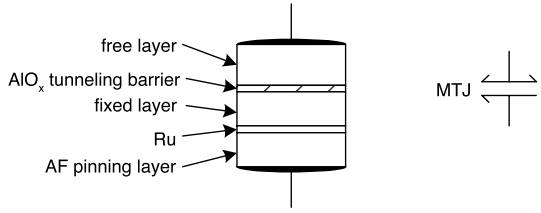


FIGURE 16.21
MTJ and circuit symbol.

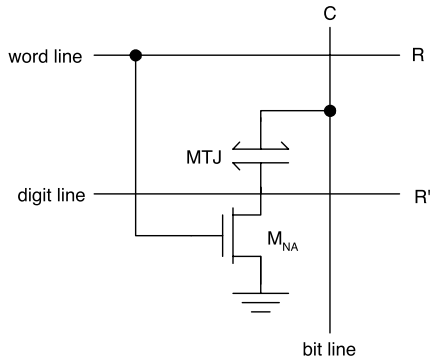


FIGURE 16.22
Magnetoresistive random access memory (MRAM) 1T1MTJ cell.

polarities fixes the magnetic field of the free layer in the antiparallel direction. Because neither current alone is sufficient to program the free layer, only one cell will be programmed, at the cross point for the bit line and the digit line. It should be noted that the digit line is not electrically connected to the MTJ; therefore, its only interaction with the MTJ is through magnetic coupling.

Readout of the stored bit is achieved by bringing the word line high to turn on the access transistor M_{NA}. Then the resistance is measured between the bit line and ground. A large difference between the resistances in the two states (~50%) makes the readout reliable and fast. Also, readout does not affect the state of the free ferromagnetic layer, so the stored bit is retained after a read operation. High density should be possible in the case of MRAM because of the simple cell design. Each bit requires only one transistor and one magnetic tunnel junction (1T1MTJ cell). This is similar to the cases for DRAM and FRAM (both requiring 1T1C cells). MRAM access times should be similar to those for the DRAM or FRAM, but the MRAM does not require refreshing after reading. This could result in a speed advantage for MRAM.

Ovonic unified memory (OUM) uses a chalcogenide alloy such as GeSbTe — the same type of material used in CD R/W and DVD R/W technology. The chalcogenide material can exist indefinitely in one of two phases: crystalline or amorphous. The amorphous phase is characterized by high resistivity,

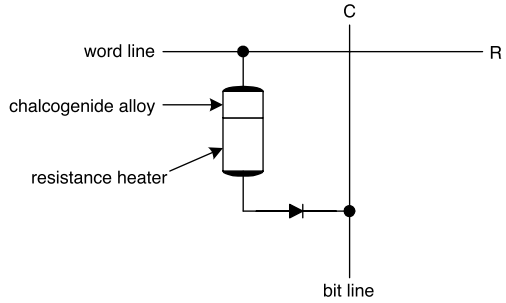


FIGURE 16.23
OUM cell.

TABLE 16.1

Comparison of Emerging Memory Technologies to Flash Memory

Parameter	Flash	FRAM	MRAM	OUM
Maximum capacity (Mb)	256	64	1	4
Cell size[a]	1	2	1.5	0.7
Erase/write cycles	10^6	10^{16}	10^{14}	10^{12}
Read/write voltages (V)	2/12	1.5/1.5	3.3/3.3	0.4/1
Read/write speed (ns)	20/1000	40/40	50/50	50/50

[a] Normalized to the cell size for flash memory.

whereas the crystalline phase exhibits low resistivity. It is therefore possible to use one phase to represent logic one and the other to represent logic zero.* The ratio of the two resistivities is 100, making read operations reliable and fast. The chalcogenide alloy can be made amorphous by heating it above its melting temperature and then allowing it to cool rapidly. The crystalline phase can be realized by heating the material to slightly below its melting temperature, thus allowing it to crystallize by a process of *solid phase epitaxy*. Each cell of the OUM comprises a programmable chalcogenide resistor, a resistance heater, and an isolation diode, as shown in Figure 16.23.

Some of the properties of these emerging memory technologies are summarized in Table 16.1. Although none of these memories can match the capacity of DRAM or flash memory, FRAM has entered the commercial marketplace and MRAM will follow soon. It is likely that these technologies will coexist in the marketplace with DRAM and flash memory for some time, unless significant advances give one particular technology a decisive advantage.

* CD R/W and DVD R/W technology utilizes the dramatic difference in optical reflectivity for the two phases of the chalcogenide.

16.11 Summary

Digital memories store bits of information in cells that are arranged in rectangular arrays. An array with 2^N rows and 2^M columns requires N + M address bits and provides 2^{N+M} cells. Each cell may contain one or more bits. If each cell contains L bits, then the memory chip will have L data lines. The data in any individual cell may be accessed by selecting the j*th* row and the k*th* column. The row is selected by applying an N-bit row address, which is decoded by the row decoder, and the column is selected by applying an M-bit address, which is decoded by the column decoder. The column decoding circuitry usually performs double duty, transferring data in and out of the memory chip. Therefore, bits of data are transferred on column lines within the memory core. For this reason, the column lines are called bit lines while the row lines are called word lines.

In large digital systems, data storage is organized according to capacity and speed. The highest capacity media (which are also the slowest) are farthest from the processor. These are nonvolatile storage media such as magnetic and optical disk drives; next are nonvolatile memory circuits and, still closer to the processor, are volatile memory circuits. Finally, additional memory circuitry is placed on the processor chip itself. This memory has limited capacity but is optimized for speed of access.

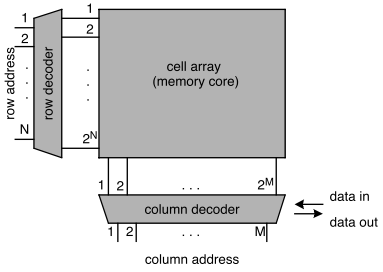
Memory chips are classified as read-only memory (ROM) or random access memory (RAM). Data can be written to or read from RAM. Random access memories can be further classified as static RAM (SRAM) and dynamic RAM (DRAM). Because static RAMs store information in latches, these chips retain their data as long as the system power is on, without the need for clocking or refreshing. Dynamic RAMs store information using charges on capacitors. Because these capacitors exhibit some level of charge leakage, the voltages must be sensed and refreshed every few milliseconds to prevent data loss. SRAMs and DRAMs are inherently volatile in nature.

Read-only memories can be classified according to their capabilities for programming and erasing. Those circuits called simply "ROM" are factory programmed and may not be erased or reprogrammed after fabrication. Programmable read-only memory (PROM) may be programmed by the customer one time only; no provision is made for erasure or reprogramming. Erasable programmable read-only memory (EPROM) may be programmed, erased, and reprogrammed many times; however, erasure requires removing the chip from the system for flood exposure by ultraviolet radiation. Electrically erasable programmable read-only memory (EEPROM, or E^2PROM), is considerably more convenient because the erase and program operations may be done with the chip in place. Flash memory is a special type of EEPROM that allows large blocks of data to be erased quickly. All ROMs are nonvolatile.

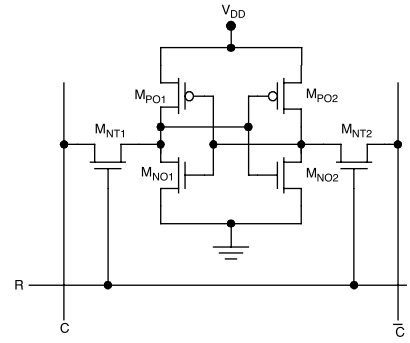
Emerging memory concepts include ferroelectric random access memory (FRAM), magnetoresistive random access memory (MRAM), and ovonic unified memory (OUM). These memories are nonvolatile and allow orders of magnitude more erase–write cycles than flash memory does.

DIGITAL MEMORIES QUICK REFERENCE

Memory Architecture

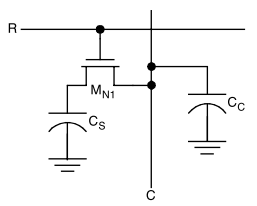


SRAM



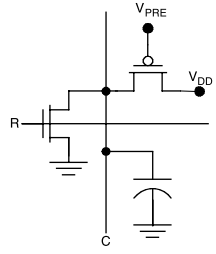
Digital memories store bits in cells arranged in rectangular arrays. Read-write memory is called random access memory (RAM) to distinguish it from read-only memory (ROM). Static RAM (SRAM) requires six transistors per bit (6T cell). Dynamic RAM (DRAM) can be realized with a 1T1C cell but must be refreshed frequently to avoid the loss of data. ROM is factory programmed but programmable ROM (PROM) is user programmed once. Electrically programmable ROM(EPROM) is programmed electrically but erased by ultraviolet exposure. Electrically erasable programmable ROM (EEPROM) may be programmed and erased in the circuit. Flash memory is similar to EEPROM but is denser and faster. Emerging nonvolatile memories such as the following:

DRAM



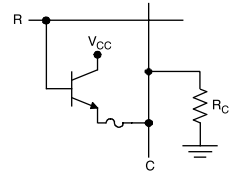
1T DRAM cell

ROM



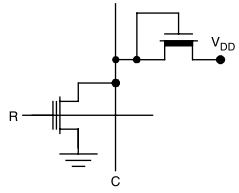
CMOS NOR ROM cell

PROM



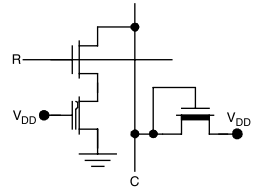
bipolar PROM cell

EPROM



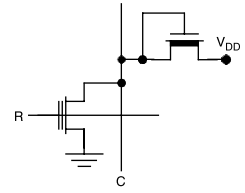
1T cell based on FAMOS transistor

EEPROM



2T cell using FLOTOX transistor

Flash Memory



1T cell based on ETOX transistor

$f = 10^{-15}$ $p = 10^{-12}$ $n = 10^{-9}$ $\mu = 10^{-6}$ $m = 10^{-3}$ $k = 10^{3}$ $M = 10^{6}$ $G = 10^{9}$

Problems

P16.1. Suppose you are designing a memory chip that will be organized with 2^N rows, 2^M columns, and L bits per address.

1. What is the required number of address and data pins in terms of N, M, and L?
2. How many pins are required for a 1-Gb memory chip if each address holds 16 bits?

P16.2. What is the minimum number of pins required for a 256-Mb memory chip? (Include V_{DD}, GND, and CLK pins.)

P16.3. Consider a 1-Mb SRAM with a square layout (1024 × 1024). The word line parasitics are 40 Ω/bit and 10 fF/bit. The column line parasitics are 1 Ω/bit and 8 fF/bit. Estimate the access time for the memory, assuming repeaters have not been used.

P16.4. Consider a 1-Mb DRAM with a square layout (1024 × 1024). The word line parasitics are 60 Ω/bit and 15 fF/bit. The column line parasitics are 1 Ω/bit and 12 fF/bit. Determine the minimum number of repeaters that must be inserted into the row lines such that the overall access time will be reduced to less than 50% of its original value.

P16.5. Consider a DRAM with 2^X bits. R_w = 65 Ω/bit and C_w = 20 fF/bit. R_b = 0.5 Ω/bit and C_b = 10 fF/bit. Determine the optimum layout (the numbers of rows and columns) such that the access time is minimized.

References

1. www.ibm.com (IBM Corporation).
2. www.micron.com (Micron Technology Inc.).
3. www.samsung.com (Samsung Corporation).
4. www.toshiba.com (Toshiba Corporation).
5. Quad data rate (QDR) SRAM design guide, Micron Technology Inc., technical note TN-54-01, www.micron.com, 2001.
6. Lage, C., Hayden, J.D., and Subramanian, C., Advanced SRAM technology — the race between 4T and 6T cells, *1996 Int. Electron. Devices Meet.*, 271, 1996.
7. Santoro, M., Tavrow, L., and Bewick, G., A subnanosecond 64 Kb BiCMOS SRAM, *Proc. 1994 Bipolar/BiCMOS Circuits Technol. Meet.*, 95, 1994.
8. Tsaur, J.J., Jih, C.W., Tsaur, H.W., and Kuo, J.B., Scaling consideration of BiCMOS SRAMs, *Proc. 1996 IEEE Int. Symp. Circuits Syst.*, 2116, 1991.
9. Dennard, R.H., Field-effect transistor memory, U.S. Patent 3,387,286, 1968.

10. Adler, E., DeBrosse, S.F. Geissler, S.F., Holmes, S.J., Jaffe, M.D., Johnson, J.B., Koburger, C.W., Lasky, J.B., Lloyd, B., Miles, G.L., Nakos, J.S., Noble, W.P., Voldman, S.H., Armacost, M., and Ferguson, R., The evolution of IBM CMOS DRAM technology, *IBM J. Res. Dev.*, 39, 167, 1995.
11. Kim, K., Hwang, C.-G., and Lee, J., DRAM technology perspective for gigabit era, *IEEE Trans. Electron. Devices*, 45, 598, 1998.
12. Noble, W. and Walker, W., Fundamental limitations on DRAM storage capacitors, *IEEE Circuits Devices*, 1, 45, 1985.
13. Kenney, D., Parries, P., Pan, P., Tonti, W., Cote, W., Dash, S., Lorenz, P., Arden, W., Mohler, R., Roehl, S., Bryant, A., Haensch, W., Hoffman, B., Levy, M., Yu, A.J., and Zeller, C., A buried-plate trench cell for 64-Mb DRAM, *Dig. Tech. Papers IEEE 1992 Symp. VLSI Technol.*, 14, 1992.
14. Nesbit, L., Alsmeier, J., Chen, J.B., DeBrosse, J., Fahey, P., Gall, M., Gambino, J., Gernhardt, S., Ishiuchi, H., Kleinhenz, R., Mandelman, J., Mii, T., Morikado, M., Nitayama, A., Parke, S., Wong, H., and Bronner, G., A 0.6-µm2 256Mb DRAM cell with self-aligned buried strap (BEST), *Tech. Dig. Papers Int. Electron. Devices Meet.*, 627, 1993.
15. Dennard, R.H.H., Scaling challenges for DRAM and microprocessors in the 21st century, *Electrochem. Soc. Proc.*, 97, 519, 1997.
16. Sunami, H., Kure, T., Hashimoto, N., Itoh, K., Toyabe, T., and Asai, S., A corrugated capacitor cell (CCC) for megabit dynamic MOS memories, *Tech. Dig. Papers 1982 Int. Electron. Devices Meet.*, 806, 1982.
17. Kang, H.K., Kim, K., Shin, Y., Park, I.S., Ko, K.M., Kim, C.G., Oh, K., Kim, S.E., Hong, C.G., Kwon, K.W., Yoo, J.Y., Kim, Y.G., Lee, C., Paick, W.S., Suh, D.I., Park, C.J., Lee, S., Ahn, S.T., Hwang, C.-G., and Lee, M., Highly manufacturable process technology for reliable 256-Mbit and 1-Gbit DRAMs, *Tech. Dig. Papers 1994 Int. Electron. Devices Meet.*, 635, 1994.
18. Bronner, G., Aochi, H., Gall, M., Gambino, J., Gernhardt, S., Hammerl, E., Ho, H., Iba, J., Ishiuchi, H., Jaso, M., Kleinhenz, R., Mii, T., Narita, M., Nesbit, L., Neumueller, W., Nitayama, A., Ohiwa, T., Parke, S., Ryan, J., Sato, T., Takato, H., and Yoshikawa, S., A fully planarized 0.25-µm CMOS technology for 256Mbit DRAM and beyond, *Dig. Tech. Papers 1995 IEEE Symp. VLSI Technol.*, 15, 1995.
19. Crowder, S., Stiffler S., Parries, P., Bronner, G., Nesbit, L., Wille, W., Powell, M., Ray, A., Chen, B., and Davari, B., Trade-offs in the integration of high-performance devices with trench capacitor DRAM, *Tech. Dig. Papers 1997 Int. Electron. Devices Meet.*, 45, 1997.
20. Crowder, S., Hannon, R., Ho, H., Sinitsky, D., Wu, S., Winstel, K., Khan, B., Stiffler, S.R., and Iyer, S.S., Integration of trench DRAM into a high-performance 0.18µm logic technology with copper BEOL, *Tech. Dig. Papers 1998 Int. Electron. Devices Meet.*, 1017, 1998.
21. Takato, H., Koike, H., Yoshida, T., and Ishiuchi, H., Embedded DRAM technology: past, present and future, *Proc. 1999 Int. Symp. VLSI Technol., Syst., Appl.*, 239, 1999.
22. Iyer, S.S. and Kalter, H.L., Embedded DRAM technology: opportunities and challenges, *IEEE Spectrum*, 36, 56, 1999.
23. Itoh, K., Nakagome, Y., Kimura, S., and Watanabe, T., Limitations and challenges of multigigabit DRAM chip design, *IEEE J. Solid-State Circuits*, 32, 624, 1997.

24. Yamagata, T., Tomishima, S., Tsukude, M., Hashizume, Y., and Arimoto, K., Circuit design techniques for low-voltage operating and/or giga-scale DRAMs, *Dig. Tech. Papers 1995 Int. Solid State Circuits Conf.*, 248, 1995.
25. Chan, T.Y., Chen, J., Ko, P.K., and Hu, C., The impact of gate-induced drain leakage current on MOSFET scaling, *Tech. Dig. Papers Int. Electron. Devices Meet.*, 719, 1987.
26. Itoh, K., Trends in megabit DRAM circuit design, *IEEE J. Solid-State Circuits*, 25, 778, 1990.
27. Hidaka, H., Fujishima, K., Matsuda, Y., Asakura, M., and Yoshihara, T., Twisted bit-line architectures for multi-megabit DRAMs, *IEEE J. Solid-State Circuits*, 24, 21, 1989.
28. Rupp, T., Chaudary, N., Dev, K., Fukuzaki, Y., Gambino, J., Ho, H., Iba, J., Ito, E., Kiewra, E., Kim, B., Maldei, M., Matsunaga, T., Ning, J., Rengarajan, R., Sudo, A., Takegawa, Y., Tobben, D., Weybright, M., Worth, G., Divakaruni, R., Srinivasan, R., Alsmeier, J., and Bronner, G., Extending trench DRAM technology to 0.15-µm groundrule and beyond, *Tech. Dig. Papers 1999 Int. Electron. Devices Meet.*, 33, 1999.
29. Mandelman, J.A., Dennard, R.H., Bronner, G.B., DeBrosse, J.K., Divakaruni, R., Li, Y., and Radens, C.J., Challenges and future directions for the scaling of dynamic random-access memory (DRAM), *IBM J. Res. Dev.*, 46, 2002.
30. Owen, W.H. and Tchon, W.E., E^2PROM product issues and technology trends, *Proc. 1989 VLSI Computer Peripherals*, 1, 1989.
31. www.siemens.com (Siemens Corporation).
32. www.amd.com (Advanced Micro Devices).
33. Barre, A.G., Flash memory — an exploding alternative to fixed hard disks, *Dig. 1993 Int. Magn. Conf.*, BZ-06, 1993.
34. Masuoka, F. and Endoh, T., Flash memories, their status and trends, *Proc. 4th Int. Conf. Solid-State IC Technol.*, 128, 1995.
35. Aritome, S., Advanced flash memory technology and trends for file storage application, *Tech. Dig., 2000 Int. Electron. Devices Meet.*, 763, 2000.
36. Wett, T. and Levy, S., Flash-the memory technology of the future that's here today, *Proc. 1995 IEEE Nat. Aerosp. Electron. Conf.*, 359, 1995.
37. Lai, S., Flash memories: where we were and where we are going, *Tech. Dig. 1998 Int. Electron. Devices Meet.*, 971, 1998.
38. Lorenzini, M., Rudan, M.V., and Baccarani, G., A dual gate flash EEPROM cell with two-bit storage capacity components, *IEEE Trans. Packag. Manuf. Technol.*, Part A, 20, 182, 1997.
39. Kynett, V., Fandrich, M.L., Anderson, J., Dix, P., Jungroth, O., Kreifels, J.A., Lodenquai, R.A., Vajdic, B., Wells, S., Winston, M.D., and Yang, L., A 90-ns one-million erase/program cycle 1-mbit flash memory, *IEEE J. Solid-State Circuits*, 24, 1259, 1989.
40. Imamiya, K., Sugiura, Y., Nakamura, H., Himeno, T., Takeuchi, K., Ikehashi, T., Kanda, K., Hosono, K., Shirota, R., Aritome, S., Shimizu, K., Hatakeyama, K., and Sakui, K., A 130 mm^2 256-Mb NAND flash with shallow trench isolation technology, *Dig. Tech. Papers 1999 IEEE Int. Solid-State Circuits Conf.*, 112, 1999.
41. www.ramtron.com (Ramtron Corporation).
42. www.fujitsu.com (Fujitsu Corporation).
43. Chung, Y., Experimental 128-kbit ferroelectric memory with 10^{12} endurance and 10-year data retention, *IEEE Proc. Circuits, Devices Syst.*, 149, 136, 2002.

44. Kim, H.H., Song, Y.J., Lee, S.Y., Joo, H.J., Jang, N.W., Jung, D.J., Park, Y.S., Park, S.O., Lee, K.M., Joo, S.H., Lee, S.W., Nam, S.D., and Kim, K., Novel integration technologies for highly manufacturable 32 Mb FRAM, *Dig. Tech. Papers 2002 Symp. VLSI Technol.*, 210, 2002.
45. Choi, M.-K., Jeon, B.-G., Jang, N., Min, B.-J., Song, Y.-J., Lee, S.-Y., Kim, H.-H., Jung, D.-J., Joo, H.-J., and Kim, K., A 0.25-μm 3.0 V 1T1C 32 Mb nonvolatile ferroelectric RAM with address transition detector (ATD) and current forcing latch sense amplifier (CFLSA) scheme, *Dig. Tech. Papers 2002 IEEE Int. Solid-State Circuits Conf.*, 162, 2002.
46. Jang, N.W., Song, Y.J., Kim, H.H., Jung, D.J., Koo, B.J., Lee, S.Y., Joo, S.H., Lee, K.M., and Kim, K., A novel 1T1C capacitor structure for high density FRAM, *Dig. Tech. Papers 2000 Symp. VLSI Technol.*, 34, 2000.
47. Lee, S.Y., Jung, D.J., Song, Y.J., Koo, B.J., Park, S.O., Cho, H.J., Oh, S.J., Hwang, D.S., Lee, S.I., Lee, J.K., Park, Y.S., Jung, I.S., and Kim, K., A FRAM technology using 1T1C and triple metal layers for high performance and high density FRAMs, *Dig. Tech. Papers 1999 Symp. VLSI Technol.*, 141, 1999.
48. Miyakawa, T., Tanaka, S., Itoh, Y., Takeuchi, Y., Ogiwara, R., Doumae, S.M., Takenakal, H., Kunishima, I., Shuto, S., Hidaka, O., Ohtsuki, S., and Tanaka, S.-I., A 0.5-μm 3 V 1T1C 1 Mb FRAM with a variable reference bitline voltage scheme using a fatigue-free reference capacitor, *Dig. Tech. Papers 1999 IEEE Int. Solid-State Circuits Conf.*, 104, 1999.
49. Kachi, T., Shoji, K., Yamashita, H., Kisu, T., Torii, K., Kumihashi, T., Fujisaki, Y., and Yokoyama, N., A scalable single-transistor/single-capacitor memory cell structure characterized by an angled-capacitor layout for megabit FeRAMs, *Dig. Tech. Papers 1998 Symp. VLSI Technol.*, 126, 1998.
50. Durlam, M., Naji, P., Omair, A., DeHerrera, M., Calder, J., Slaughter, J.M., Engel, B., Rizzo, N., Grynkewich, G., Butcher, B., Tracy, C., Smith, K., Kyler, K., Ren, J., Molla, J., Feil, B., Williams, R., and Tehrani, S., A low-power 1-Mbit MRAM based on 1T1MTJ bit cell integrated with copper interconnects, *Dig. Tech. Papers 2002 Symp. VLSI Circuits*, 158, 2002.
51. Naji, P.K., Durlam, M., Tehrani, S., Calder, J., and DeHerrera, M.F., A 256-kb 3.0-V 1T1MTJ nonvolatile magnetoresistive RAM, *Dig. Tech. Papers 2001 IEEE Int. Solid-State Circuits Conf.*, 122, 2001.
52. Tehrani, S., Durlam, M., DeHerrera, M., Chen, E., Calder, J., and Kerszykowski, G., High-density pseudo spin valve magnetoresistive RAM, *Proc. 7th Biennial IEEE Nonvolatile Memory Technol. Conf.*, 43, 1998.
53. www.ovonyx.com (Ovonyx Inc.).
54. Gill, M., Lowrey, T., and Park, J., Ovonic unified memory — a high-performance nonvolatile memory technology for stand-alone memory and embedded applications, *Dig. Tech. Papers 2002 IEEE Int. Solid-State Circuits Conf.*, 202, 2002.
55. Lai, S. and Lowrey, T., OUM—- a 180-nm nonvolatile memory cell element technology for stand-alone and embedded applications, *Tech. Dig. 2001 Int. Electron. Devices Meet.*, 36.5.1, 2001.
56. Maimon, J., Spall, E., Quinn, R., and Schnur, S., Chalcogenide-based nonvolatile memory technology, *Proc. 2001 IEEE Aerosp. Conf.*, 2289, 2001.

17

Design and Layout

17.1 Introduction

Once designed, digital integrated circuits must be transferred to a physical silicon wafer. This is done by a process of pattern transfer called lithography. The basis for lithography is a set of lithographic masks, which contain the patterns to be transferred to the semiconductor wafer. Physical design of integrated circuits involves the creation of these masks. One mask is used for each pattern transfer step in the wafer fabrication process, so more than 20 masks may be needed to fabricate a BiCMOS wafer. The many mask layers and the sheer complexity of modern VLSI circuits make physical design seem a daunting task. However, using sophisticated computer tools with libraries of devices and circuits makes the problem tractable.

17.2 Photolithography and Masks

Lithography[1–17] is the process of transferring physical patterns to the semiconductor wafer. Among the several variations on the basic lithographic process are photolithography,[1–8] x-ray lithography,[1,9–16] electron beam lithography,[1,15] ion-beam lithography,[18–22] and photoelectron lithography. The basis for all of these processes is the exposure and development of radiation-sensitive chemicals called resists.[23–27]

There are positive and negative resists. The use of a positive resist for pattern transfer is illustrated in Figure 17.1. First, a fresh layer of silicon dioxide is grown over the entire wafer. Then, the wafer is coated with a thin layer of positive photoresist.* This photoresist is spun on and then baked to the desired hardness. After baking, it is exposed to ultraviolet radiation

* Positive photoresists are available from a number of manufacturers, each with its own proprietary formulations. However, positive photoresists generally comprise the following: a low molecular weight, alkali-soluble resin (such as phenol formaldehyde novolac), a photoactive dissolution inhibitor (such as orthoquinone diazide), and a solvent (such as xylene).

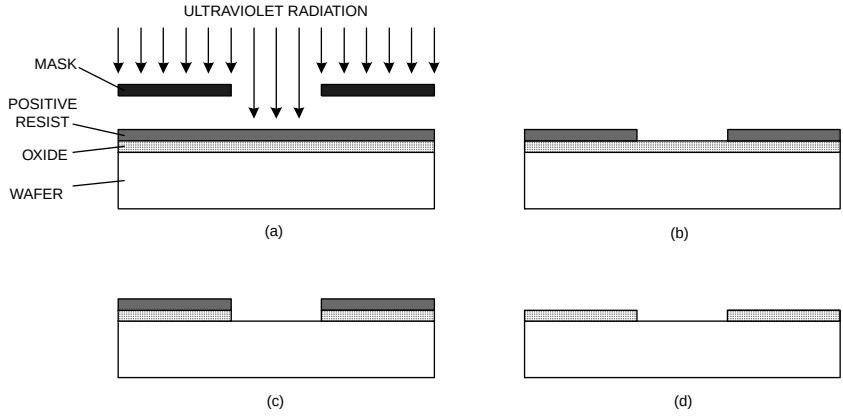


FIGURE 17.1
Pattern transfer using positive photoresist.

through a photomask. The development process involves washing in organic solvents such as acetone or xylene. Radiated areas degrade and become readily soluble in these liquids; thus, positive resists are sometimes called "degrading resists." Following development, the patterned photoresist can be used as a mask for the chemical etching of the silicon dioxide. The result is that the silicon dioxide assumes the same pattern as the original mask (a positive image) and can be used as a physical mask for diffusion, ion implantation, or etching, thus completing the pattern transfer process.

The use of negative resist is similar in many ways and is illustrated in Figure 17.2. Following the growth of a fresh layer of silicon dioxide, the negative photoresist* is spun on. After a prebake, the photoresist is exposed through a mask and hardened by a postbake. The irradiation promotes cross linking in the resist, resulting in high molecular weight chains difficult to remove. For this reason, negative resists are occasionally called "cross-linking resists." Upon development, only the unexposed resist is removed. The remaining resist forms a mask for patterning the underlying oxide layer. The transferred pattern is the same as with positive resist, but the mask must be the negative image of the desired pattern.

A problem with negative resist is the solvent-induced swelling that occurs during development, which results in ragged edges and poor resolution. For this reason, positive resist is capable of higher resolution and is used exclusively for VLSI today. The photomasks are designed using computer tools and the resulting designs reside in computer files. The designs are transferred to masks** using electron beam (e-beam) lithography in which an e-beam is steered directly by the computer using a raster-scan or vector-scan approach.

---

* Negative resist comprises a synthetic rubber (such as cyclized *cis*-polyisoprene) with a radiation-sensitive cross-linking agent (such as bisazide) in an organic solvent base.

** Photomasks are typically made using quartz substrates and metal mask layers (such as chromium or Fe_2O_3).

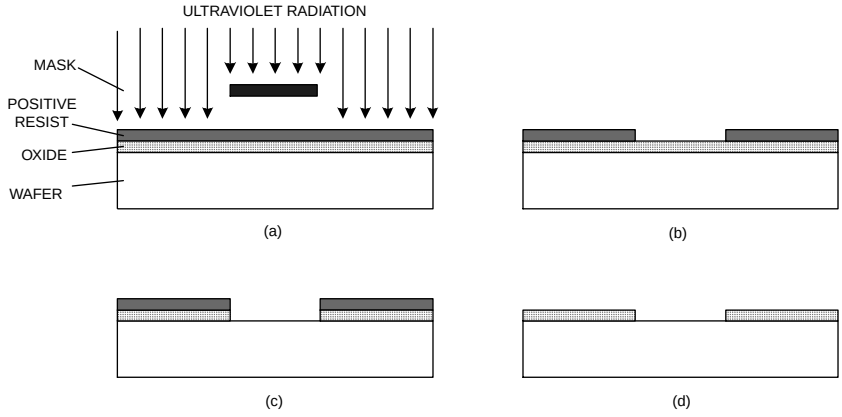


FIGURE 17.2
Pattern transfer using negative resist.

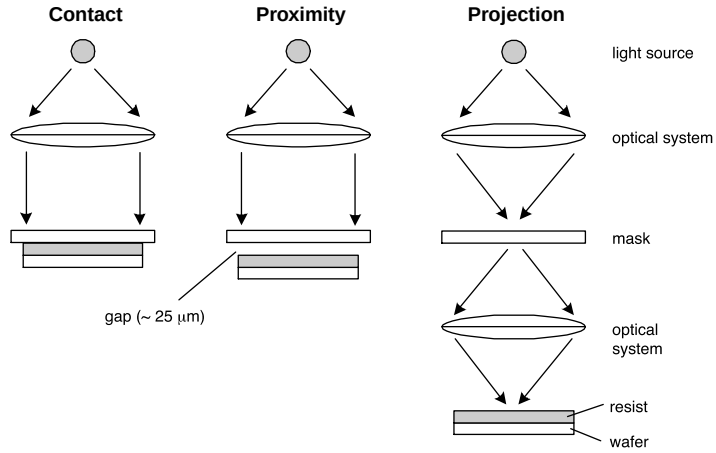


FIGURE 17.3
Contact, proximity, and projection photolithographic printing systems.

Pattern transfer from the masks to the wafers is done by ultraviolet photolithography because of its higher throughput compared to e-beam lithography. This printing can be done using a contact, proximity, or projection approach as shown in Figure 17.3; however, step-and-repeat (S/R) projection printing is used in all modern VLSI fabrication lines.[28–30] This is because S/R projection printing provides extended mask life compared to contact printing and better resolution than proximity printing.

The basic limitations of optical lithography are related to the optical wavelength. The minimum feature size is determined by the diffraction limit and is given by

$$2X = \frac{\lambda}{2NA}, \tag{17.1}$$

where λ is the optical wavelength and NA is the numerical aperture. The depth of focus for the optical system is

$$d = \frac{\lambda}{(NA)^2}. \tag{17.2}$$

Steady reduction of the minimum feature size from one technology generation to the next has mandated the reduction of the optical wavelength — despite the use of phase contrast masks,[31] which have allowed resolution performance exceeding the diffraction "limit." At the present time, deep ultraviolet (DUV) lithography systems are in use and extreme ultraviolet (EUV) systems[32–34] are being readied for deployment. For a given optical wavelength, the choice of numerical aperture thus involves a trade-off between the resolution and depth of focus.

17.3 Layout and Design Rules

Design rules specify the minimum dimensions and spacings that may be used in a layout design. These values are related to the fabrication process as well as the optical wavelength used for printing. Design rules may be scalable or absolute. Scalable rules are stated in terms of *X* (where the minimum feature size is 2X)*; absolute design rules are stated in terms of microns. Scalable rules have the advantage that they can be applied to different process lines with different values of *X*. However, they may not be simultaneously optimized for different values of *X* because some design rules do not scale with *X*, so worst-case values must be used to produce a scalable set of design rules. In practice, scalable and absolute design rules are used today. An example of a scalable design rule set is that used by the VLSI prototyping service MOSIS.[35,36] This scalable design rule set may be downloaded from the MOSIS Web site.[35]

Generally speaking, the three types of design rules are 1) minimum widths, 2) minimum spacings, and 3) minimum surrounds. Specific examples of these classes of rules will be discussed in the following sections. Throughout this chapter, discussions will focus on layout principles and will be kept as general as possible. For the sake of specific examples, however, a scalable n-well CMOS process will be assumed. A state-of-the-art CMOS process was

* Often, the minimum feature size is denoted 2λ. Here, the notation 2X has been used to avoid confusion with the optical wavelength used for photolithography.

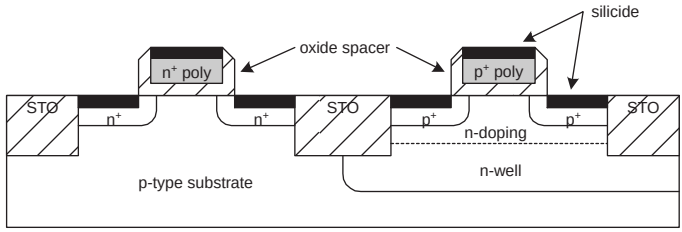


FIGURE 17.4
Complementary MOSFETs made by a scalable n-well CMOS process.

TABLE 17.1

Scalable n-Well CMOS Layers[a]

physical layer	name	layout symbol
n well	NWELL	
silicon nitride	ACTIVE	
polysilicon	POLY1	
p+ implant	PSELECT	
n+ implant	NSELECT	
contact cut	CONTACT	
metal 1	METAL1	
metal 2	METAL2	

[a] Modern CMOS processes use two layers of polysilicon and eight or more layers of metal; however, the principles of layout design may be illustrated without invoking this level of complexity.

outlined in Chapter 1 and allows the fabrication of complementary MOSFETs as shown in Figure 17.4. The basic layers and layout legends for such a scalable n-well CMOS process are summarized in Table 17.1.

Note that the ACTIVE layer defines the placement of silicon nitride. In turn, this silicon nitride is used to pattern shallow trench oxide (STO); the STO is grown wherever the nitride is *absent*. Therefore, channel regions are defined by the overlap of the active and polysilicon layers. A single mask is used to pattern the polysilicon wires — even though these wires exist with p-type and n-type doping — because the polysilicon is doped simultaneously with the source and drain regions of the MOSFETs. (This is required by the *self-aligned process*.)

17.3.1 Minimum Linewidths and Spacings

The minimum linewidth 2X is the smallest dimension permitted for any feature in the layout; 2X is also called the "minimum feature size." Technologically, the minimum feature size corresponds to the minimum width for a polysilicon line. For example, with 0.1-μm technology, the minimum polysilicon linewidth is 0.1 μm and the value of X is 0.05 μm.

The minimum linewidths and spacings are determined primarily by the process technology and equipment used, especially the wavelength used for the photolithography. However, they are also determined in part by lateral doping and depletion effects. Implanted regions spread laterally during the annealing process, resulting in lateral doping; the diffusion of impurities also results in lateral doping effects. In addition, depletion regions surround implantations or diffusions made in a semiconductor of opposite conductivity type. The lateral doping and the depletion regions affect the minimum spacings of doped regions. Violation of the minimum linewidth or spacing rules may result in a nonfunctioning circuit because of broken lines (if the minimum linewidth is violated) or a short circuit (if the minimum spacing between lines is violated).

The design rules for polysilicon, stated in terms of X, are illustrated in Figure 17.5; that is, the minimum linewidth for polysilicon is 2X and the minimum spacing for two polysilicon lines is 3X. The design rules for implantations are illustrated in Figure 17.6 and Figure 17.7. The minimum width for implanted regions is greater than for polysilicon to allow for depletion effects at the edges of the doped region. The minimum spacing design rule for implanted regions of opposite conductivity has been made large to avoid the latch-up problem discussed in Chapter 9. (Two important exceptions to the rule are shown in Figure 17.7: first, a PSELECT region can be abutted by an NSELECT region used to contact the NWELL and, second, an NSELECT region may be abutted by a PSELECT region used to contact the substrate.)

The spacing and width design rules for the metal 1 and metal 2 layers are illustrated in Figure 17.8 and Figure 17.9. To allow for registration errors between mask levels, the METAL1 design rules are more conservative than the POLYSILICON design rules, and the METAL2 design rules are more conservative than the METAL1 rules.

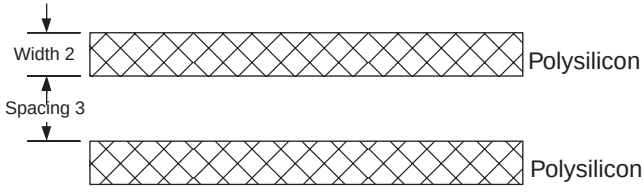


FIGURE 17.5
Polysilicon design rules.

FIGURE 17.6
Design rules for two implantations of the same type.

FIGURE 17.7
Design rules for implantations of opposite conductivity type.

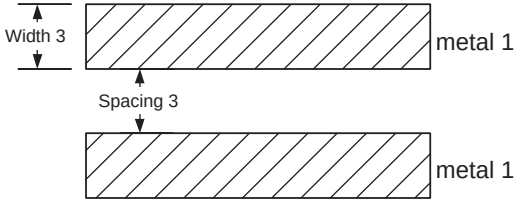


FIGURE 17.8
Metal 1 design rules.

FIGURE 17.9
Metal 2 design rules.

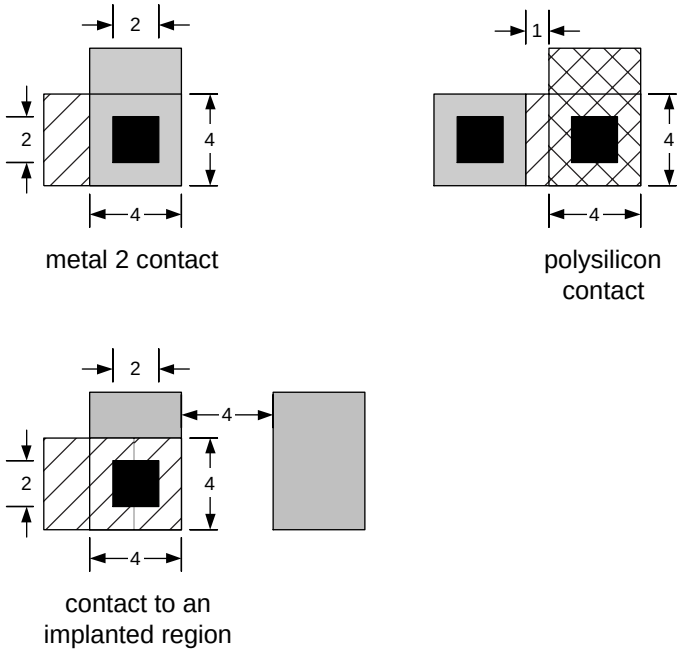


FIGURE 17.10
Layout design rules for contacts.

17.3.2 Contacts and Vias

Contacts are made to n+, p+, or polysilicon device regions by opening windows in the overlying oxide prior to metallization. For a scalable rule set, the minimum dimension for a metal contact is 2X; in practice, all contact cuts are made this size. Therefore, an increase in contact area is achieved using multiple contact cuts, rather than a single, large area cut.

The minimum surround for a metal contact is X which means that the layer contacted must extend one half the minimum feature size in all directions. This allows for tolerance in registration between the two mask levels. Contact cuts must also be made in the upper glassy layers placed between the levels of metal. An example is the case of a contact made between metal 1 and metal 2. Such cuts are often called vias. However, their layout rules are the same as for other contact cuts. The basic design rules for contacts are illustrated in Figure 17.10.

17.3.3 MOSFETs

The basic design rules for a MOSFET are illustrated in Figure 17.11 for the case of a p-channel MOSFET. The channel is formed where the polysilicon wire overlaps the p select implant. The minimum width for this polysilicon

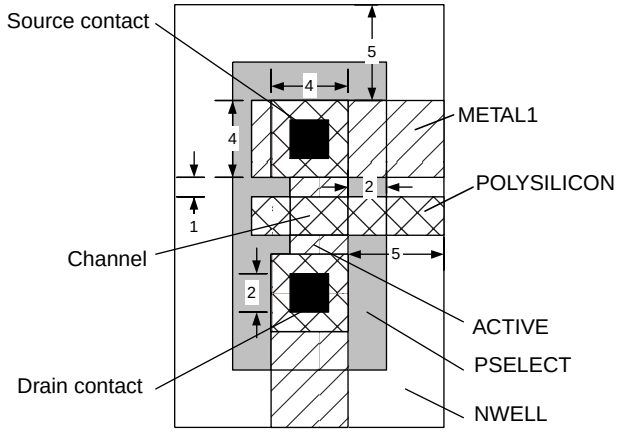


FIGURE 17.11
MOSFET design rules (p-MOSFET).

wire is 2X. Also, the polysilicon wire must extend beyond the active region by at least 2X on either side. The metal contacts must be 2X on a side and the p select region must extend 2X in all directions around the contact cuts. The p select contact regions must be spaced at least 1X from the channel. The PSELECT region must extend beyond the ACTIVE region by 2X in all directions and the NWELL must extend 5X beyond the ACTIVE region by 5X in all directions. Notice that the total area of the transistor scales with X^2 so that halving the minimum feature size will reduce the transistor area by a factor of one quarter.

Often it is necessary to connect MOSFETs in series. One such situation is illustrated in Figure 17.12 for the case of two series-connected p-channel MOSFETs. In such a case it is not necessary to form source and drain contact regions between the series-connected MOSFETs. Instead, the common PSELECT region between the two channels forms the drain of one transistor and the source of the other. The minimum separation between the two channel regions is 2X. Also, outside the channels, the minimum separation between a polysilicon wire and the PSELECT region is 1X. Polysilicon wires may overlap or cross the METAL1 layer because a glassy insulator layer exists between them. N-MOSFET design rules are similar and illustrated in Figure 17.13. Compared to the p-MOSFET, a key difference is that the n-MOSFET does not require a well (if n-well technology is utilized).

17.3.4 Bipolar Transistors

Whereas the n-well scalable CMOS process is optimized for the fabrication of MOSFETs, it does allow the design and fabrication of npn bipolar transistors also. The layout of such a bipolar transistor is shown in Figure 17.14. The NSELECT layer is used for the emitter and the ohmic contact to the collector; the base is fabricated using the PSELECT layer and the collector

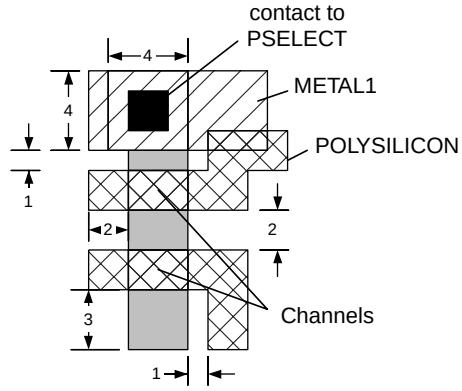


FIGURE 17.12
Design rules for series connected MOSFETs (p-MOSFETs).

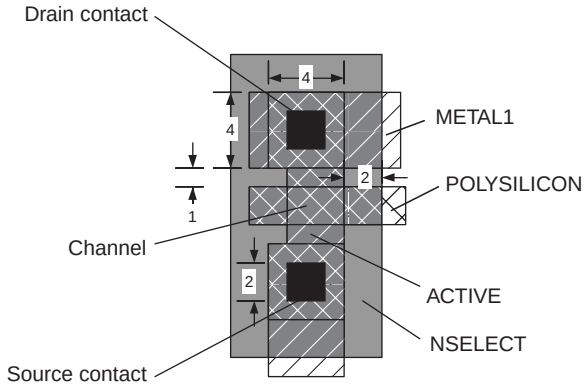


FIGURE 17.13
N-MOSFET design rules.

is an NWELL layer. Multiple contacts are made to the emitter, base, and collector to reduce the parasitic contact resistances. The entire device is isolated by the surrounding p-type semiconductor, sometimes using a surrounding shallow trench oxide as well.

All contacts are designed to be 2X on a side. The base, emitter, and collector regions must extend 2X beyond the contacts in all directions. The base must extend 2X beyond the emitter region. The NWELL collector must extend 6X beyond the base and collector contact regions in all directions. The npn bipolar transistor design illustrated in Figure 17.14 represents a nonoptimized device compromised for the sake of compatibility with the CMOS process. The layout of high-performance bipolar transistors is qualitatively similar; however, n+ subcollectors, polysilicon emitters, and shallow trench oxide are used to reduce parasitics and improve the current gain cutoff frequency.

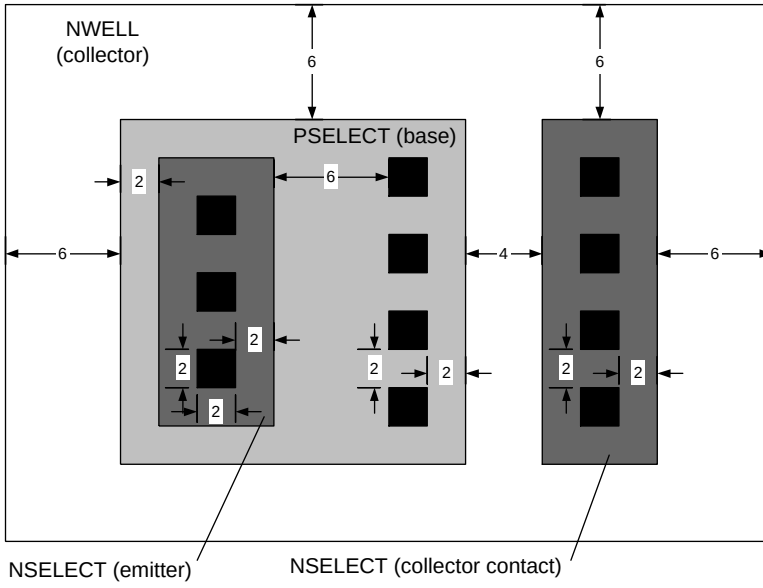


FIGURE 17.14
Layout design rules for an npn bipolar transistor.

17.3.5 Resistors

For the case of a scalable n-well CMOS technology, resistors are made using an implanted p-type semiconductor (PSELECT) surrounded by an n-type region (NWELL) that provides isolation. Typically, such resistors are made using a minimum-width PSELECT region, the length of which is designed to achieve the desired resistance. The PSELECT region may wind back and forth several or many times as shown in Figure 17.15, depending on the required resistance. The value of an integrated resistor may be determined by counting the number of squares in the resistor and multiplying by the sheet resistance for the PSELECT layer. The end bells and corners are counted as half squares.

Example 17.1

Determine the value of the integrated resistor shown in Figure 17.16, assuming that the sheet resistance of the PSELECT layer is 200 Ω/square.

Solution. Taking into account the two end bells and four corners, the value of the resistor is

$$R = 200\,\Omega/\text{square}\left(\frac{26}{3}+\frac{3.5}{3}+\frac{27}{3}+\frac{3.5}{3}+\frac{26}{3}+\frac{4\,\text{corners}}{2}+\frac{2\,\text{end bells}}{2}\right)$$

$$= 6.3\ \text{k}\Omega$$

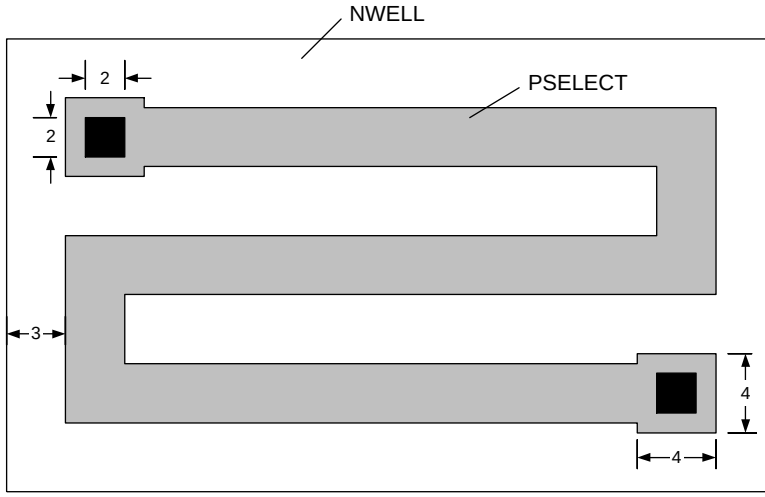


FIGURE 17.15
Design rules for a resistor made by a p-type implantation into an n well.

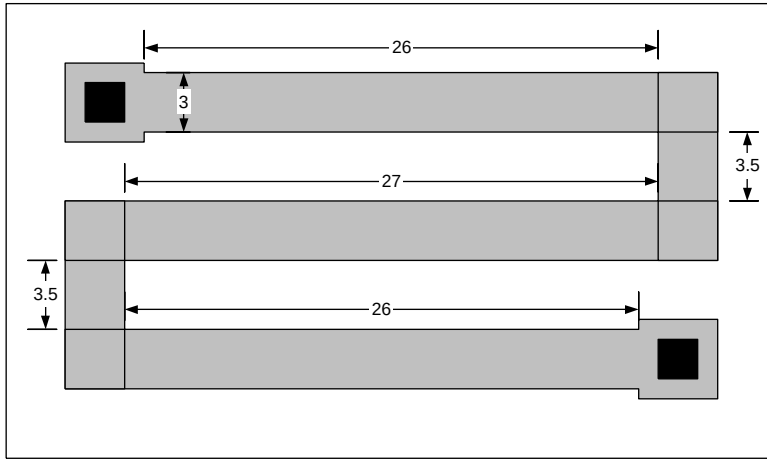


FIGURE 17.16
Example resistor for determination of the resistance.

17.4 Physical Design of CMOS Circuits

The physical design of CMOS gates involves bringing together the circuit concepts of Chapter 9 with the layout rules introduced in this chapter. The simplest example is that of the inverter.

Example 17.2

Create the physical layout for a minimum-size inverter and determine the required chip area in terms of X.

Solution. The minimum channel length is equal to 2X, the minimum linewidth for polysilicon, and the minimum channel width is 3X, the minimum width for implanted regions. Minimum-size transistors are designed with channel dimensions 4X × 2X to facilitate contacts to the implanted region. A minimum-size inverter uses n-channel and p-channel devices that are minimum-size transistors. A possible layout design for the minimum-size inverter is illustrated in Figure 17.17. Notice that the PSELECT region for the p-MOSFET is spaced by 10X from the NSELECT region for the n-MOSFET. This spacing eliminates the possibility of latch-up.

The NWELL surrounding the p-channel device must be connected to the most positive voltage in the circuit (V_{DD}); an NSELECT layer and metal contact are used for this purpose. As explained earlier, the NSELECT region for the NWELL contact is allowed to abut the PSELECT region. The p-type substrate must be connected to the most negative voltage in the circuit (ground); a PSELECT layer and metal contact are made to achieve this. This PSELECT region is allowed to abut the NSELECT region of the n-MOSFET.

The smallest rectangle that can enclose the minimum-size inverter laid out as shown is 51X × 18X, corresponding to an area of $918X^2$. In the case of 0.25-μm technology, approximately 7 million such inverters would fit in an area 1 cm × 1 cm. It is important to note that this minimum-size inverter does not have matched transistors. The n-channel device has a greater transconductance parameter than the p-channel device. Consequently, the DC and transient characteristics will not be symmetric.

Example 17.3

Create the physical layout for a minimum-size symmetric inverter (with matched transistors) and determine the required chip area in terms of X.

Solution. The minimum-size n-channel MOSFET has channel dimensions of 4X × 2X. The p-channel device must be scaled up by a factor of 2.5, resulting in channel dimensions of 10X × 2X. A possible layout design for the minimum-size symmetric inverter is illustrated in Figure 17.18. The scaled-up p-channel device uses two contacts each for the source and drain. The smallest rectangle that can enclose the minimum-size inverter is 39X × 25X, corresponding to an area of $975X^2$. For the case of 0.25-μm technology, 6.5 million such inverters would fit on a 1 cm × 1 cm die.

17.5 VLSI Design Principles

Modern VLSI circuits often contain more than 10^7 transistors. This level of complexity poses a significant challenge, but computer layout tools[37–41] make

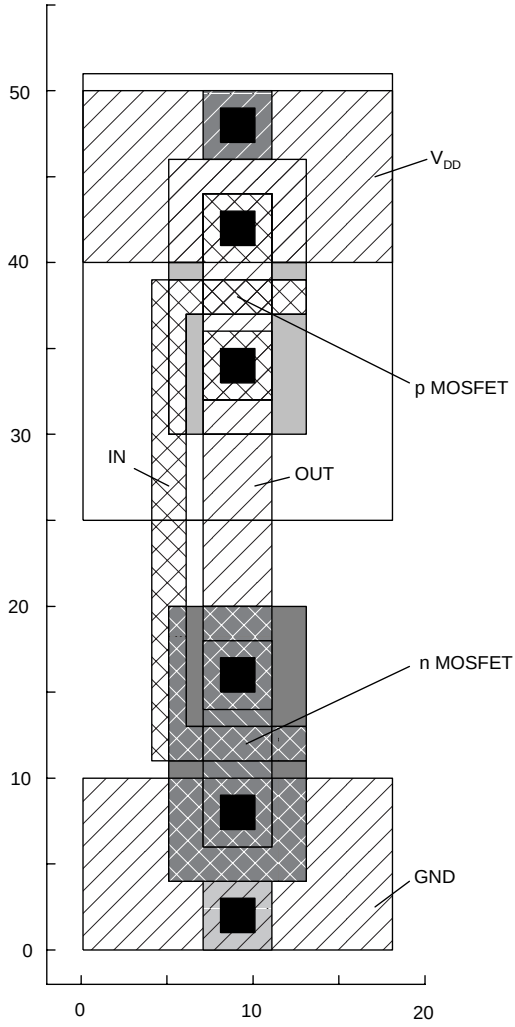


FIGURE 17.17
Minimum-size CMOS inverter.

the problem tractable. The three basic design philosophies for VLSI are 1) fully custom, 2) standard cell based, and 3) gate array based.

In a fully custom design,[42] every single gate in the integrated circuit must be designed in full detail. The advantage of this approach is that the performance and chip areas can be optimized (at least in principle). The fully custom approach requires the greatest effort, and therefore time, to arrive at the finished design.

Modern design tools contain libraries of devices, gates, and other "standard cells" that can be exploited in a cell-based design. Using standard cells[42–49]

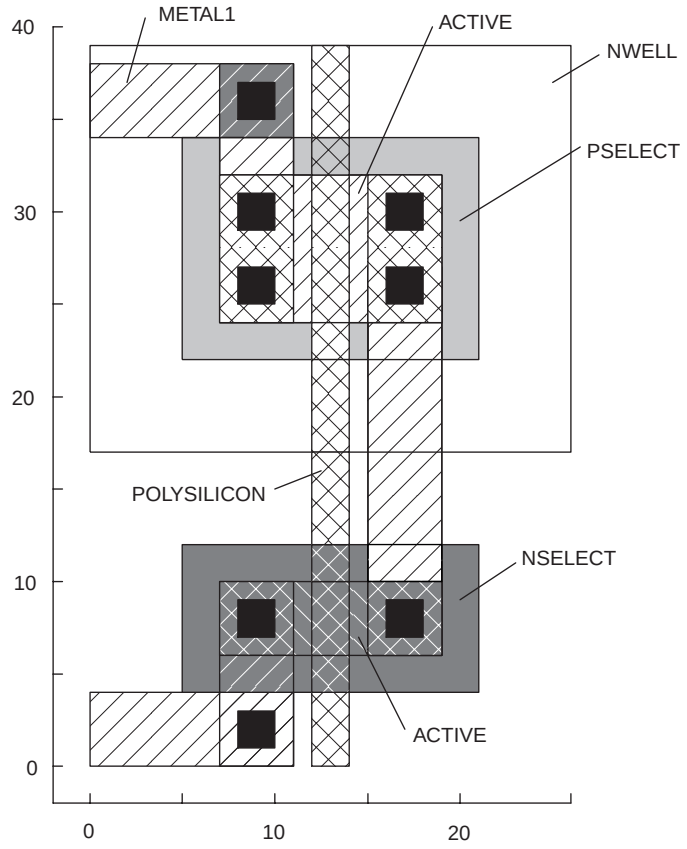


FIGURE 17.18
Minimum-size symmetric CMOS inverter.

greatly simplifies the design effort and shortens the time to market, although this efficiency comes with some penalties in performance and packing density. Nonetheless, this trade-off is often worthwhile, e.g., with application-specific integrated circuits (ASICs), for which time to market is critical in determining the success of a particular product.

Gate arrays contain regular arrays of MOSFETs, which can be connected to provide the desired logic functions. Minimal design effort is required because only the transistor interconnections must be configured; rapid prototyping is therefore possible. However, the compromises in packing density and performance make gate arrays unsuitable for mass-produced, high-density, or performance-critical designs. A variation of the gate array approach is the field programmable gate array (FPGA), which is programmed by the customer *after* fabrication.

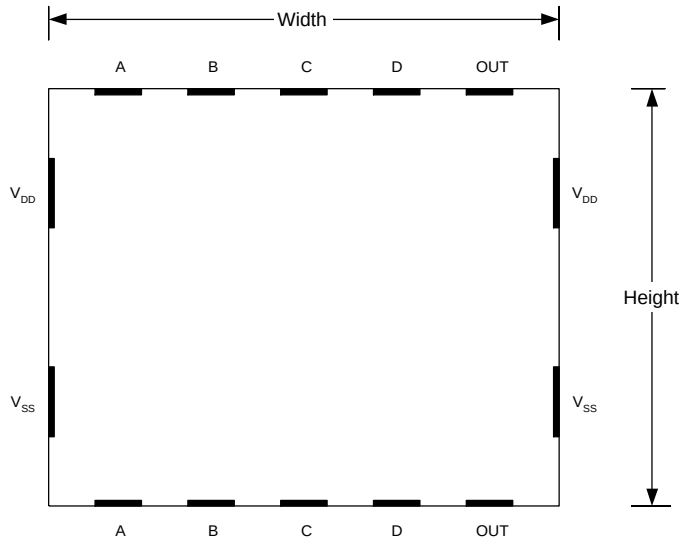


FIGURE 17.19
Basic layout of a standard cell.

Most VLSI designs involve some combination of the custom and cell-based approaches. Thus, individual gates are produced using standard cells, whereas larger subsystems, signal routing, and power routing are customized. This hybrid approach shortens the total design time with little compromise in performance.

Standard cells are rectangular; the basic layout is shown in Figure 17.19. Power connections are made to the sides of the cell whereas input/output connections are made to the top and bottom. Such standard cells are all made with uniform height so that they may be placed in rows, as shown in Figure 17.20. The design of the placement of these standard cells, called floor planning, should be done in such a way as to minimize the interconnect lengths. This is a combinatorial optimization problem, so the optimum solution may only be found by trying all possible solutions. In practice, pseudo optimization can be achieved with a reduction in computation time but only a modest compromise in performance.

Whereas standard cells are suitable for the realization of simple logic functions, *macro cells* are utilized for more complex functions such as arithmetic units and memories. For the case of macro cells, the standard height restriction is lifted, allowing greater design flexibility, although this prevents placing macro cells in rows and complicates placement and routing problems.

A special difficulty arises with respect to routing the clock signal. The clock must be routed over the entire chip area as the power rails are. Unlike the power rails, the clock lines carry a time-varying signal. Phase differences across a large area die can thus cause timing problems. This difficulty, called

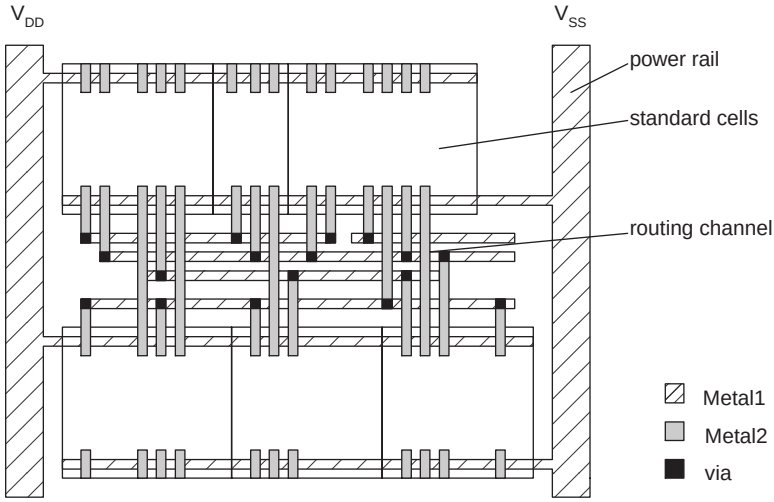


FIGURE 17.20
Standard cell placement and routing.

clock skew, is increasingly important as chip complexity and size continue to increase. One current line of research is directed at elimination of the clock for *asynchronous* integrated circuits. This approach has the advantages of eliminating clock skew and causing all circuits to run at their highest speed, at which they are limited by their own propagation delays rather than the clock.

Example 17.4

Design the layout for a CMOS standard cell realizing the NAND4 function. Assume that the standard cell inverter uses n-channel devices with dimensions 4X × 2X and p-channel devices with channel dimensions of 10X × 2X.

Solution. The circuit diagram for the CMOS circuit is shown in Figure 17.21. The electrical path from the output to ground involves four series n-MOSFETs, so these devices must be scaled by a factor of four compared to the inverter. Therefore, the n-MOSFETs will use gate dimensions of 16X × 2X. Every path from V_{DD} to the output involves a single p-MOSFET; thus, the p-MOSFETs need not be scaled and will use channel dimensions of 10X × 2X. One possible standard cell design is shown in Figure 17.22. Of course, the actual design may be influenced by restrictions on the cell height and placement of the V_{DD} and GND connections.

Example 17.5

Design the layout for a CMOS standard cell that realizes the function $Y = \overline{AB + C}$. Assume that the standard cell inverter uses n-channel devices with

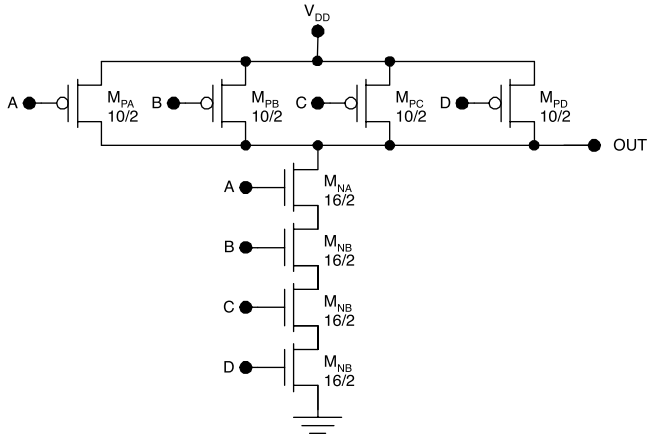


FIGURE 17.21
CMOS circuit to realize the NAND4 function.

dimensions 2X × 4X and p-channel devices with channel dimensions of 2X × 10X.

Solution. The circuit diagram for the CMOS circuit is shown in Figure 17.23. The worst-case path from the output to ground involves two series n-MOSFETs, so these devices must be scaled by a factor of two compared to the inverter. Therefore, the n-MOSFETs will use gate dimensions of 8X × 2X. The worst-case path from V_{DD} to the output involves two series p-MOSFETs; thus, the p-MOSFETs must also be scaled by a factor of two, resulting in channel dimensions of 20X × 2X. One possible standard cell design is shown in Figure 17.24.

Example 17.6

Design the layout for a CMOS standard cell that realizes the function $Y = \overline{AB + CD}$. Assume that the standard cell inverter uses n-channel devices with dimensions 4X × 2X and p-channel devices with channel dimensions of 10X × 2X.

Solution. The circuit diagram for the CMOS circuit is depicted in Figure 17.25. The worst-case path from OUT to ground involves two series n-MOSFETs, so these devices must be scaled by a factor of two compared to the inverter. Therefore, the n-MOSFETs will use gate dimensions of 8X × 2X. The worst-case path from V_{DD} to OUT involves two series p-MOSFETs; thus, the p-MOSFETs must also be scaled by a factor of two, resulting in channel dimensions of 20X × 2X. A standard cell design is shown in Figure 17.26.

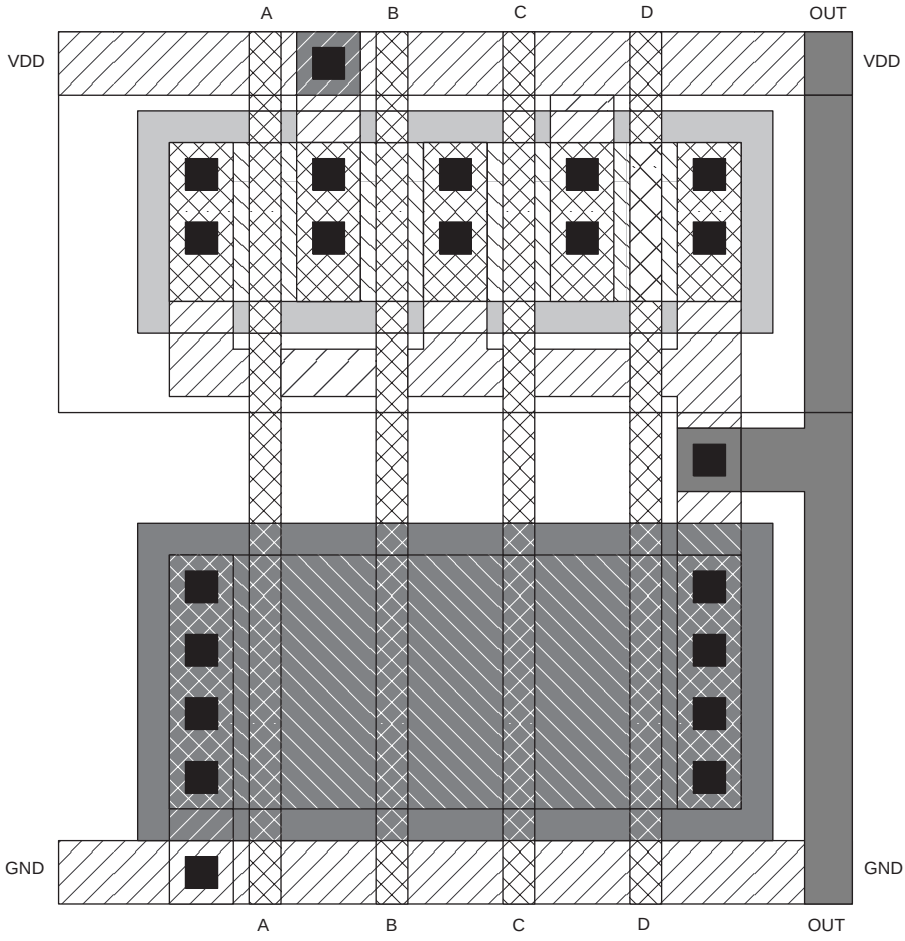


FIGURE 17.22
Layout design of a CMOS standard cell that realizes the NAND4 function.

17.6 Summary

Integrated circuits are fabricated by a sequence of steps that transfer the desired patterns to doped regions, oxide, metals, and other deposited films. This process of pattern transfer is done by lithography with a set of lithographic masks. Currently, the pattern transfer method involves using deep ultraviolet radiation and is called photolithography. The pattern transfer process therefore involves exposing a photosensitive chemical (photoresist)

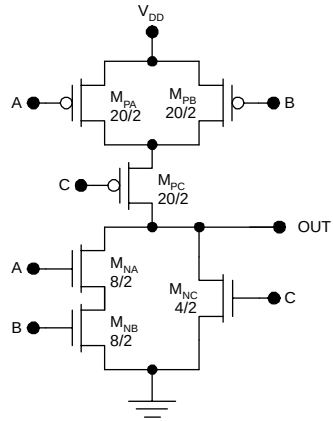


FIGURE 17.23
CMOS circuit to realize the function $Y = \overline{AB + C}$.

through the mask and then developing it. The patterned resist can then be used to etch the underlying material in the desired pattern. VLSI design is the process of laying out the patterns to be produced on the lithographic masks.

Layout design must follow a set of basic design rules, including minimum linewidths, minimum spacings, and minimum surrounds. In a scalable rule set, all design rules are given in terms of the minimum linewidth 2X (the minimum width for a polysilicon wire). In an absolute rule set, all design rules are given in microns. Scalable rule sets have the advantage that they can be easily used on different process lines with different minimum linewidths; however, they cannot be optimized for all minimum linewidths. Therefore, an absolute design rule set may allow somewhat more efficient use of chip area. In either case, the minimum linewidths are chosen to avoid breaks and open circuits, whereas the minimum spacings and surrounds are chosen to avoid misalignment problems, short circuits, or problems of latch-up.

The design rules depend on the process technology used. For the case of a scalable n-well CMOS technology such as that outlined in Chapter 1, the mask layers that must be designed include the n–well (NWELL), the silicon nitride that defines the thick field oxide (ACTIVE), p implantation (PSELECT), n implantation (NSELECT), polysilicon (POLY), contact cuts (CONTACT), first layer of metal (METAL1), and second layer of metal (METAL2). One set of scalable design rules (used by MOSIS, the integrated circuit prototyping service) for an n-well CMOS process is described in this chapter. Often, additional layers of polysilicon and metal are used, requiring additional masks designed with the same rules as the METAL2 layer.

VLSI circuits often contain millions of transistors. This level of complexity requires use of sophisticated computer tools such as those of Cadence,

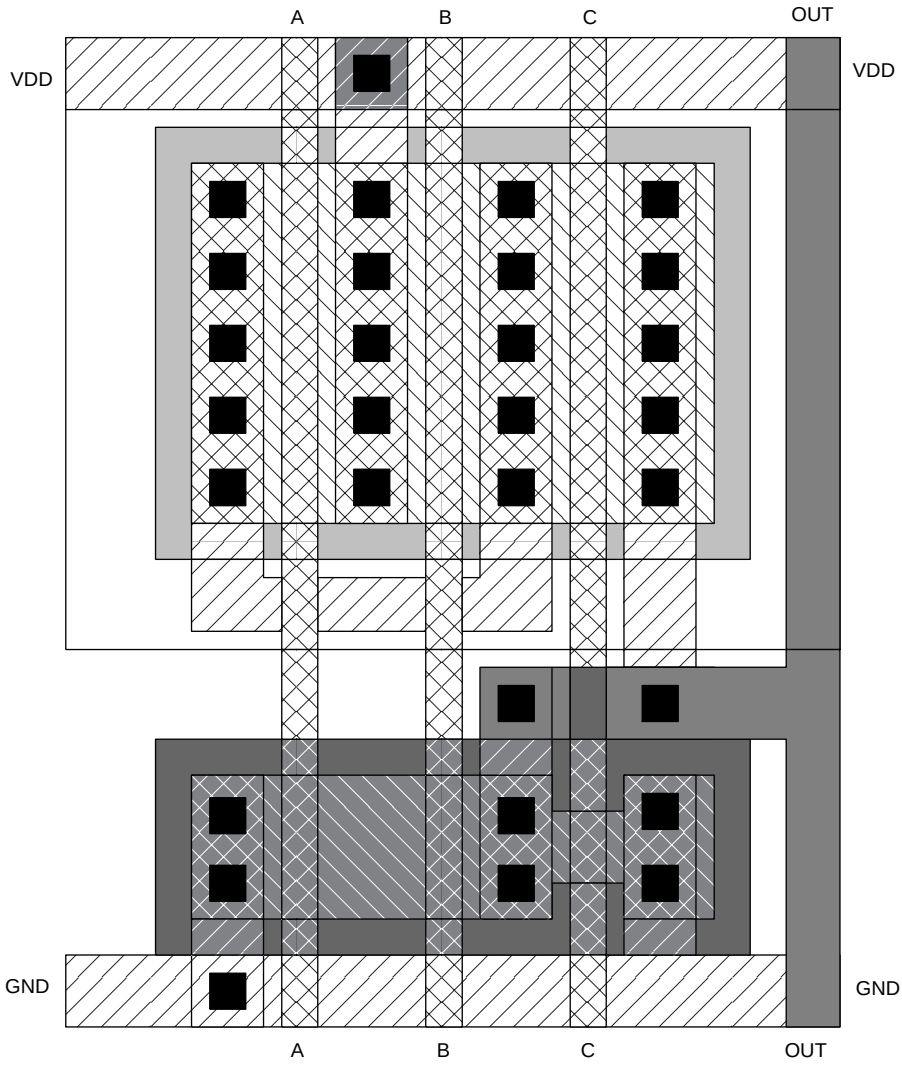


FIGURE 17.24
Layout design of CMOS standard cell that realizes the function $Y = \overline{AB + C}$.

Mentor Graphics, Synopsys, or Avant! to make the layout design tractable. Three basic design philosophies are 1) fully custom, 2) cell based, and 3) gate array based. The fully custom approach allows the highest performance and most efficient use of chip area, whereas gate array-based design allows the shortest turnaround. In practice, most designs utilize a hybrid approach involving use of some standard cells and some custom design.

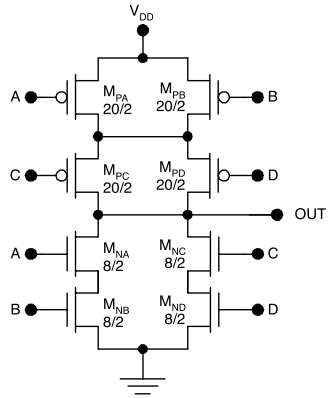


FIGURE 17.25
CMOS circuit to realize the function $Y = \overline{AB + CD}$.

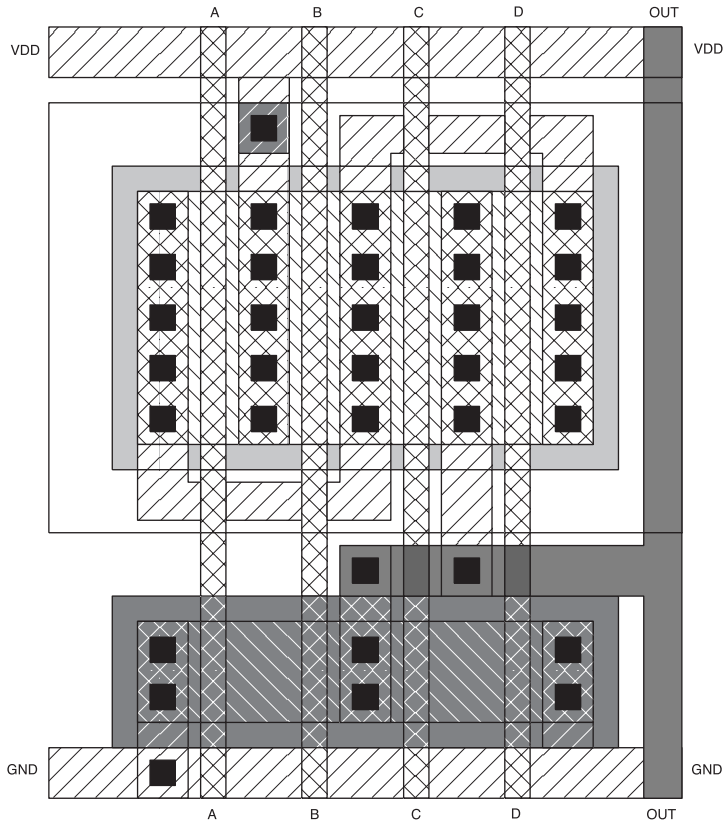
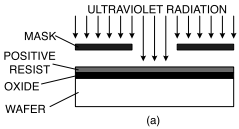
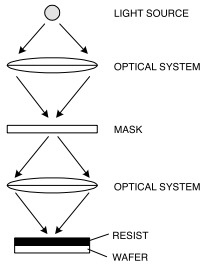


FIGURE 17.26
Layout design of CMOS standard cell that realizes the function $Y = \overline{AB + CD}$.

DESIGN AND LAYOUT QUICK REFERENCE

Photolithography and Masks





$$2X = \frac{\lambda}{2NA}$$

$$d = \frac{\lambda}{(NA)^2}$$

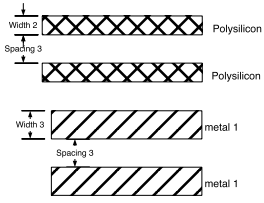
Physical design of integrated circuits involves the layout of the doped regions, oxide, contact cuts, metals, and other layers necessary to form the circuit components. These patterns are transferred to the wafer using masks and projection photolithography. Scalable design rules are stated in terms of X, where 2X is the minimum feature size. Computer tools are used to make the layout design process tractable for complex circuits and systems.

Layout Design Rules

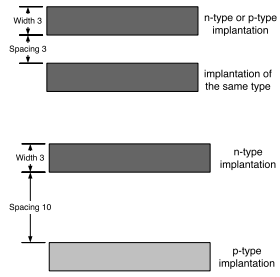
Scalable CMOS layers

physical layer	name	layout symbol
n well	NWELL	
silicon nitride	ACTIVE	
polysilicon	POLY1	
p+ implant	PSELECT	
n+ implant	NSELECT	
contact cut	CONTACT	
metal 1	METAL1	
metal 2	METAL2	

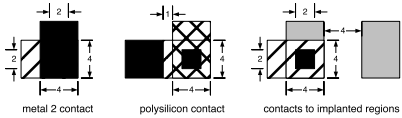
Polysilicon and Metal



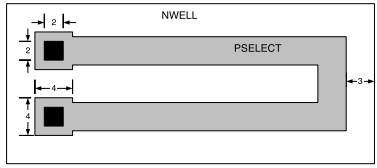
Implantations



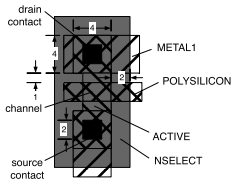
Contacts



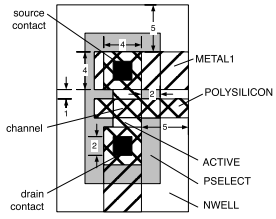
Resistors



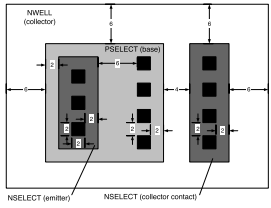
n-MOSFETs



p-MOSFETs



npn Bipolar transistors



$f = 10^{-15}$ $p = 10^{-12}$ $n = 10^{-9}$ $\mu = 10^{-6}$ $m = 10^{-3}$ $k = 10^{3}$ $M = 10^{6}$ $G = 10^{9}$

Problems

P17.1. Determine the value of the integrated resistor illustrated in Figure 17.27, assuming that the sheet resistance of the PSELECT layer is 200 Ω/square.

P17.2. Small-valued resistors can be fabricated as shown in Figure 17.28. Estimate the value of the resistor shown, assuming it behaves as four parallel resistors. The sheet rho of the PSELECT layer is 200 Ω/square.

P17.3. Lay out a 2-kΩ resistor using a PSELECT layer with a sheet rho of 200 Ω/square.

P17.4. Lay out a 50-kΩ resistor using a PSELECT layer with a sheet rho of 250 Ω/square.

P17.5. Design and lay out an n-MOSFET with a device transconductance parameter of 2 mA/V^2. Assume a 1-μm n-well CMOS technology using an oxide thickness of 10 nm.

P17.6. Design and lay out an n-MOSFET with a device transconductance parameter of 50 μA/V^2. Assume a 1-μm n-well CMOS technology using an oxide thickness of 10 nm.

P17.7. Design and lay out a p-MOSFET with a device transconductance parameter of 2 mA/V^2. Assume a 1-μm n-well CMOS technology using an oxide thickness of 10 nm.

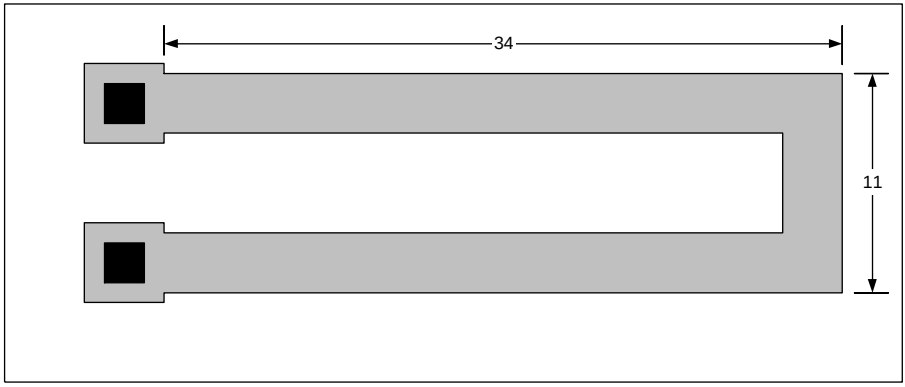


FIGURE 17.27
Integrated resistor (P17.1).

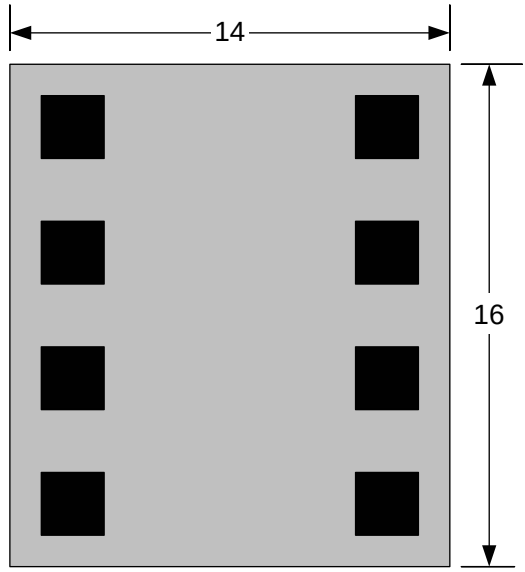


FIGURE 17.28
Small-valued resistor (P17.2).

P17.8. Consider a symmetric eight-input CMOS NOR gate, compatible with inverters having p-MOSFET gate dimensions of 10X × 2X and n-MOSFET gate dimensions of 4X × 2X. Determine the required gate dimensions for p-MOSFETs and n-MOSFETs in the eight-input NOR gate.

P17.9. Consider a symmetric eight-input CMOS NAND gate, compatible with inverters having minimum-size n-MOSFETs. Determine the required gate dimensions for p-MOSFETs and n-MOSFETs in the eight-input NAND gate.

P17.10. Lay out a symmetric four-input CMOS NAND gate using p-MOSFET gate dimensions of 10X × 2X and n-MOSFET gate dimensions of 16X × 2X. Determine the required chip area.

P17.11. Lay out a symmetric four-input CMOS NOR gate compatible with inverters having minimum-size n-MOSFETs.

P17.12. Design the layout for a CMOS standard cell that realizes the function $Y = \overline{AB + C + DE}$.

P17.13. Design the layout for a CMOS standard cell that realizes the function $Y = A \oplus B$.

P17.14. Design the layout for a CMOS standard cell one-bit full adder.

References

1. Fukuda, H. and Okazaki, S., Analysis of critical dimension control for optical-, EB-, and x-ray lithography below the 0.2-μm region, *Dig. Tech. Papers 1995 Symp. VLSI Technology*, 77, 1995.
2. Fritze, M., Chen, C.K., Astolfi, D.K., Yost, D.R., Burns, J.A., Chen, C-L., Gouker, P.M., Suntharalingam, V., Wyatt, P.W., and Keast, C.L., Enhanced resolution for future fabrication, *IEEE Circuits Devices Mag.*, 19, 43, 2003.
3. Harriott, L.R., Limits of lithography, *Proc. IEEE*, 89, 366, 2001.
4. Brunner, T., Pushing the limits of lithography for IC production, *Tech. Dig. 1997 Int. Electron. Devices Meet.*, 9, 1997.
5. Van den Hove, L., Goethals, A.M., Ronse, K., Van Bavel, M., and Vandenberghe, G., Lithography for sub-90-nm applications, *Tech. Dig. 2002 Int. Electron. Devices Meet.*, 3, 2002.
6. Matsuo, T., Endo, M., Kishimura, S., Misaka, A., and Sasago, M., Lithography solution for 65-nm node system LSIs, *Dig. Tech. Papers 2002 Symp. VLSI Technol.*, 196, 2002.
7. Pugh, G., Canning, J., and Roman, B., Impact of high-resolution lithography on IC mask design, *Proc. 1998 IEEE Custom IC Conf.*, 149, 1998.
8. Zacharias, A., X-ray lithography for integrated circuit development and manufacturing, *IEEE Trans. Components, Hybrids, Manuf. Technol.*, 5, 118, 1982.
9. Murphy, J.B., X-ray lithography sources: a review, *Proc. 1989 Particle Accelerator Conf.*, 2, 757, 1989.
10. Maldonado, J.R., Overview of x-ray lithography at IBM using a compact storage ring, *Conf. Rec. 1991 IEEE Particle Accelerator Conf.*, 542, 1991.
11. Longo, R., Chaloux, S., Chen, A., Krasnoperova, A., Lee, S., Murphy, G., Thomas, A., Wasik, C., Weybright, M., and Bronner, G., An evaluation of x-ray lithography using a 0.175-μm (0.245-μm^2 cell area) 1-Gb DRAM technology, *Dig. Tech. Papers 1998 Symp. VLSI Technol.*, 82, 1998.
12. Nakayama, Y., Recent progress and future developments in EB mask writing for x-ray lithography, *Dig. Papers 1999 Int. Microprocesses NanoTechnol. Conf.*, 8, 1999.
13. Uchiyama, S., Current status and issues of x-ray masks, *Proc. 1998 Int. Conf. Microelectronic Test Struct.*, 61, 1998.
14. Mizusawa, N., Uda, K., Tanaka, Y., Ohta, H., and Watanabe, Y., Technology and performance of x-ray stepper for volume production, *Dig. Papers 2000 Int. Microprocesses NanoTechnol. Conf.*, 108, 2000.
15. Fukuda, M. and Taguchi, T., Performance of x-ray stepper for next-generation lithography, *Dig. Papers 1999 Microprocesses NanoTechnol. Conf.*, 10, 1999.
16. Harriott, L.R., SCALPEL: projection electron beam lithography, *Proc. 1999 Particle Accelerator Conf.*, 595, 1999.
17. Melngailis, J., Focused ion beam lithography and implantation, *Proc. 8th Univ./Gov./Ind. Microelectron. Symp.*, 70, 1989.
18. Kim, Y.S., Hong, W., Woo, H.J., Choi, H.W., Kim, K.D., and Lee, S., Ion beam lithography using membrane masks, *Dig. Papers 2001 Int. Microprocesses NanoTechnol. Conf.*, 148, 2001.

19. Buchmann, L.-M., Schnakenberg, U., Torkler, M., Loschner, H., Stengl, G., Traher, C., Fallmann, W., Stangl, G., and Cekan, E., Lithography with high depth of focus by an ion projection system, *Proc. 1992 IEEE Micro Electro Mech. Syst*., 67, 1992.
20. Paek, S.W., Park, S.-H., Lee, H.Y., and Chung, H.B., Sub-0.1-μm patterning characteristics of inorganic resists by focused-ion-beam lithography, *Proc. Int. Microprocesses NanoTechnol. Conf*., 129, 1998.
21. Melngailis, J., Ion sources for nanofabrication and high resolution lithography, *Proc. 2001 Particle Accelerator Conf*., 76, 2001.
22. www.shipley.com (Shipley Company).
23. www.dupont.com (Dupont Corporation).
24. Sarantopoulou, E., Cefalas, A.C., Gogolides, E., and Argitis, P., Photoresist polymeric materials for 157-nm photolithography, *Dig. 2000 Eur. Conf. Lasers Electro-Opt*., 1, 2000.
25. Kishimura, S., Endo, M., and Sasago, M., High-performance 157-nm resist based on fluorine-containing polymer, *Dig. Tech. Papers 2001 Symp. VLSI Technol*., 37, 2001.
26. Kishimura, S., Sasago, M., Shirai, M., and Tsunooka, M., Approach of various polymers to 157-nm single-layer resists, *Proc. Int. Microprocesses NanoTechnol. Conf*., 104, 2000.
27. www.asml.com (ASML Holding).
28. www.canon.com (Canon Incorporated).
29. www.nikon.com (Nikon Corporation).
30. Misaka, A., Matsuo, T., and Sasago, M., Super-resolution enhancement method with phase-shifting mask available for random patterns, *Dig. Tech. Papers 2002 Symp. VLSI Technol*., 200, 2002.
31. Stulen, R.H. and Sweeney, D.W., Extreme ultraviolet lithography, *IEEE J. Quantum Electron*., 35, 694, 1999.
32. Gwyn, C.W., Stulen, R.H., Sweeney, D.W., and Attwood, D.T., Extreme ultraviolet lithography, *J. Vac. Sci. Technol. B*, 16, 3142, 1998.
33. Owa, S., Shiraishi, N., Omura, Y., Aoki, T., Matsumoto, Y., Hatasawa, M., Mori, T., and Tanaka, I., Development of F2 exposure tools, *Proc. 2001 Int. Microprocesses NanoTechnol. Conf*., 308, 2001.
34. www.mosis.org (MOSIS).
35. Pina, C., Low cost IC prototyping from the MOSIS Service, *Proc. 1999 IEEE Int. Conf. Microelectron. Syst. Educ*., 1, 1999.
36. www.research.digital.com/wrl/projects/magic/magic.html ("Magic" CAD layout tool).
37. www.cadence.com (Cadence CAD layout tool).
38. www.mentor.com (Mentor Graphics CAD layout tool).
39. www.synopsys.com (Synopsys CAD layout tool).
40. www.avanticorp.com (Avant! CAD layout tool).
41. Eriksson, H., Larsson–Edefors, P., Henriksson, T., and Svensson, C., Full-custom vs. standard-cell design flow an adder case study, *Proc. 2003 Asia S. Pac. Design Automation Conf*., 507, 2003.
42. Koike, K., Kawai, K., Onozawa, A., Takei, Y., Kobayashi, Y., and Ichino, H., High-speed, low-power, bipolar standard cell design methodology for Gbit/s signal processing, *IEEE J. Solid-State Circuits*, 33, 1536, 1998.
43. Lin, S., Marek–Sadowska, M., and Kuh, E.S., Delay and area optimization in standard-cell design, *Proc. 27th ACM/IEEE Design Automation Conf*., 349, 1990.

44. Ramachandran, K., Cordell, R.R., Daly, D.F., Deutsch, D.N., and Kwan, A.F., SYMCELL — a symbolic standard cell design system, *Proc. 1990 IEEE Custom IC Conf.*, 16.1/1, 1990; Integrated Circuits Conference, 1990.
45. Vygen, J., Algorithms for detailed placement of standard cells, *Proc. 1998 Design, Automation Test Eur.*, 321, 1998.
46. Prasad, R.K. and Koren, I., The effect of placement on yield for standard cell designs, *Proc. 2000 IEEE Int. Symp. Defect Fault Tolerance VLSI Syst.*, 3, 2000.
47. Cho, K. and Song, M., Design of novel macro-cells for next generation ASIC cell library, *Proc. 7th Int. Conf. Electron., Circuits Syst.*, 320, 2000. 2000. ICECS 2000.
48. Dash, R.K., Pramod, T., Vasudevan, V., and Ramakrishna, M., A transistor level placement tool for custom cell generation, *Proc. 13th Int. Conf. VLSI Design*, 254, 2000.
49. Shanbhag, A., Danda, S., and Sherwani, N., Floorplanning for mixed macro block and standard cell designs, *Proc. 4th Great Lakes Symp. VLSI*, 26, 1994.
50. Upton, M., Samii, K., and Sugiyama, S., Integrated placement for mixed macro cell and standard cell designs, *Proc. 27th ACM/IEEE Design Automation Conf.*, 32, 1990.

18

Integrated Circuit Packages

18.1 Introduction

Once digital integrated circuits have been designed and fabricated on a wafer, the wafer is cut into rectangular die,* which are tested and packaged for assembly in systems. Packaging requirements for VLSI circuits are rather stringent, requiring large numbers of electrical connections, capability of high input and output data rates, and efficient removal of large quantities of heat. Moreover, these packages must be compact, lightweight, inexpensive, and reliable. Entire books have been written on this important subject. The intent of this chapter is not to go into such detail, but rather to provide a sound introduction to the principles involved.

18.2 Package Types

There are five basic types of integrated circuit packages[1]:

- **Through-hole packages** have metal pins that may be inserted through holes drilled in the circuit board for soldering. Through-hole technology (THT) has been around the longest, but is inefficient in its utilization of printed circuit area.
- **Surface mount technology (SMT) packages** utilize metal leads that can be soldered to a single surface of the printed circuit board. They are much smaller and more lightweight than through-hole packages for a given number of electrical connections; in addition, they are more resistant to mechanical shock. Surface mount packages are growing in popularity for these reasons. In fact, some product applications would not have been possible without surface mount components; these include laptop computers, PDAs, digital wireless phones, and digital camcorders, to name a few.

* The plural of die is dice; however, it is standard practice in industry to use "die" as the plural.

- **Chip-scale packages** represent the most compact packaging scheme apart from the use of bare die; the package dimensions typically are only 20% greater than the die dimensions. On the other hand, chip-scale packages offer advantages in handling and testability compared to bare die. Usually, they are attached to circuit boards via an array of metal bumps. This technology provides a high pin density and is mechanically robust.
- **Bare die, or *unpackaged* parts**, offer the minimum size and weight and also eliminate RC time delays associated with the package leads. The significant challenges associated with this technology include handling, testing, mounting, and reliability.
- **Module assemblies** combine bare die or, occasionally, packaged die in a module. They introduce another level of packaging between the integrated circuit and the circuit board; surface mount and through-hole modules are used in practice. Some modules use stacked die to achieve the minimum connection lengths and the highest efficiency in circuit board utilization.

Today an almost endless variety of integrated circuit packages is available. Some standards* have been established (for example, by the Joint Electron Device Engineering Council, or JEDEC); however, manufacturers are introducing new packages at an ever increasing rate, some of which are unique to a single product or product line. Therefore, no attempt will be made to catalog them all. Instead, the basic concepts behind package designs will be presented with some important examples. The reader is referred to manufacturers' Web sites for up-to-date information on package types.

18.2.1 Through-Hole Packages

Through-hole packages[1–9] have metal pins that may be inserted through holes drilled in the circuit board for soldering. Dual in-line packages (DIPs), quad in-line packages (QIPs), and pin grid arrays (PGAs) are three main types of through-hole packages. DIPs are rectangular packages with metal pins arranged along two sides; an example is illustrated in Figure 18.1. QIPs have pins arranged along all four sides of the package for higher efficiency. PGAs utilize pins arranged in a rectangular grid on the bottom of the package and can be designed to accommodate a relatively large number of electrical connections. A 68-pin PGA is shown in Figure 18.2.

DIPs are by far the most popular THT packages and come in a number of varieties. Plastic DIPs (PDIPs) are the most cost effective, whereas ceramic DIPs (CERDIPs) are more suitable for high-power, high-temperature applications. Shrink DIPs (SDIPs, also known as skinny DIPs or SK-DIPs) utilize

* Unfortunately, many integrated circuit package standards are based on the English system of units for historical reasons. Thus, pin spacings are sometimes specified in mils (1000 mil = 1 in.).

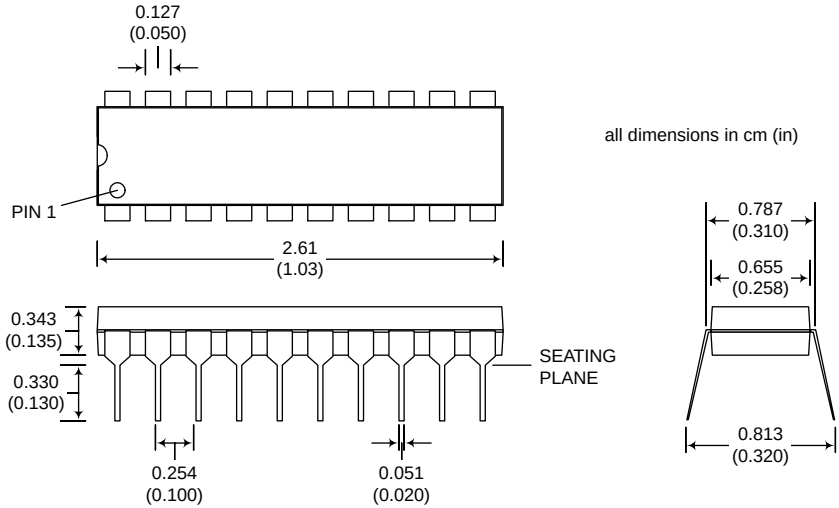


FIGURE 18.1
Plastic dual in-line package (PDIP) with 20 pins.

closer lead spacing and are more compact. Zig-zag in-line packages (ZIPs) achieve even closer lead spacings in two zig-zag patterns. Quad in-line packages (QIPs, also known as QUIPs) utilize leads on all four sides. This advantage is slight compared to shrink dips and is offset by greater difficulty in handling. PGAs are superior to the other THT packages in terms of pin efficiency and heat removal. Plastic and ceramic versions are available.

18.2.2 Surface Mount Packages

Surface mount packages[1–5,10–14] are compact, lightweight, and mechanically robust. Inexpensive applications use molded plastic, which greatly simplifies the manufacturing process. The plastic is simply molded over the metal lead frame. However, this process brings the plastic in direct contact with the die so that the thermal expansion mismatch is an issue. Hermetically sealed ceramic and metal surface mount packages are also available and avoid this problem.

Surface mount packages include small outline integrated circuits (SOICs), quad flat packs (QFPs), J-leaded chip carriers (LCCs), and ball grid array (BGA) packages. SOICs have gull wing leads that are soldered to the top surface of the circuit board. Quad flat packs are similar to SOICs but have leads on all four sides. J-lead chip carriers have J-shaped leads that bend under the package; they may be surface mounted or socketed. BGA packages use a grid of bottom-mounted solder balls for attachment to the circuit board. Of these SMT packages, the most popular are variations of the BGA, SOIC, and LCC (such as the plastic J-lead chip carrier, or PLCC). Several important types of surface mount packages are depicted in Figure 18.3 through Figure 18.6.

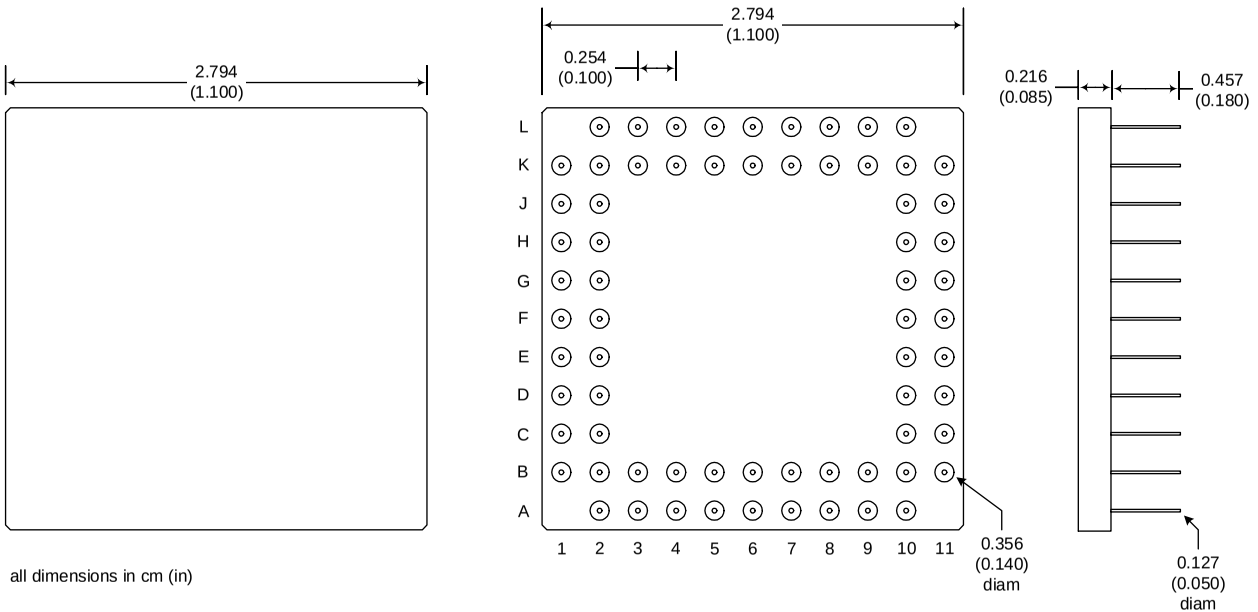


FIGURE 18.2
Plastic pin grid array (PGA) package with 68 pins.

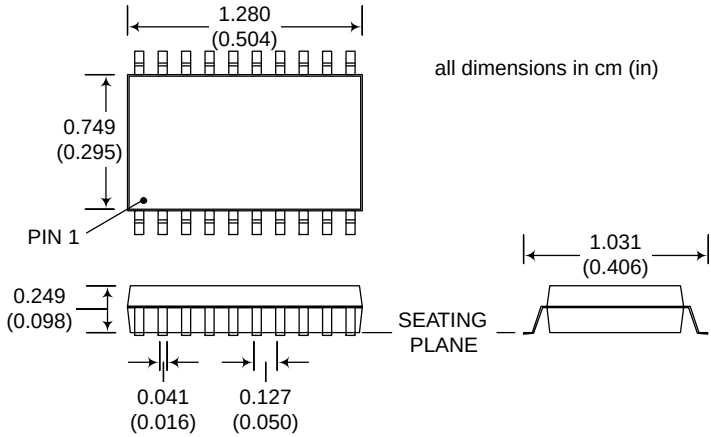


FIGURE 18.3
Small outline integrated circuit (SOIC) package with 20 pins.

18.2.3 Chip-Scale Packages

Chip-scale packages[1–5,15–20] are designed to be only slightly (<20%) larger than the die they house. On the other hand, they provide benefits in ease of handling and testability compared to bare die. Chip-scale packaging technologies include the popular micro ball grid array (μBGA) package styles. Nearly all chip-scale packages utilize flip-chip technology; thus, the die is mounted top down on a ceramic substrate. Prior to die mounting, the aluminum pads on the die are built up with metal bumps that form one-to-one attachments to a pattern of metal pads (the *land*) on the substrate. In turn, the substrate is attached to a circuit board or module using an array of solder bumps.

18.2.4 Bare Die

Unpackaged die[1–5,21–26] offer minimum size and weight; they also eliminate signal delays associated with the package. *Chip on board* (COB) technology involves bonding the die directly on the circuit board, face up, followed by wire bonding. Bare die may also be mounted directly on the circuit board by a flip-chip approach using solder balls. A third approach involves using bare die mounted on polyamide film with metal traces on it (*tape automated bonding*, or TAB). An example of this approach is *chip on flex* (COF).

18.2.5 Multichip Modules

Multichip modules (MCMs)[27–34] may use through-hole or surface mount technology; the feature that distinguishes them is the placement of more than

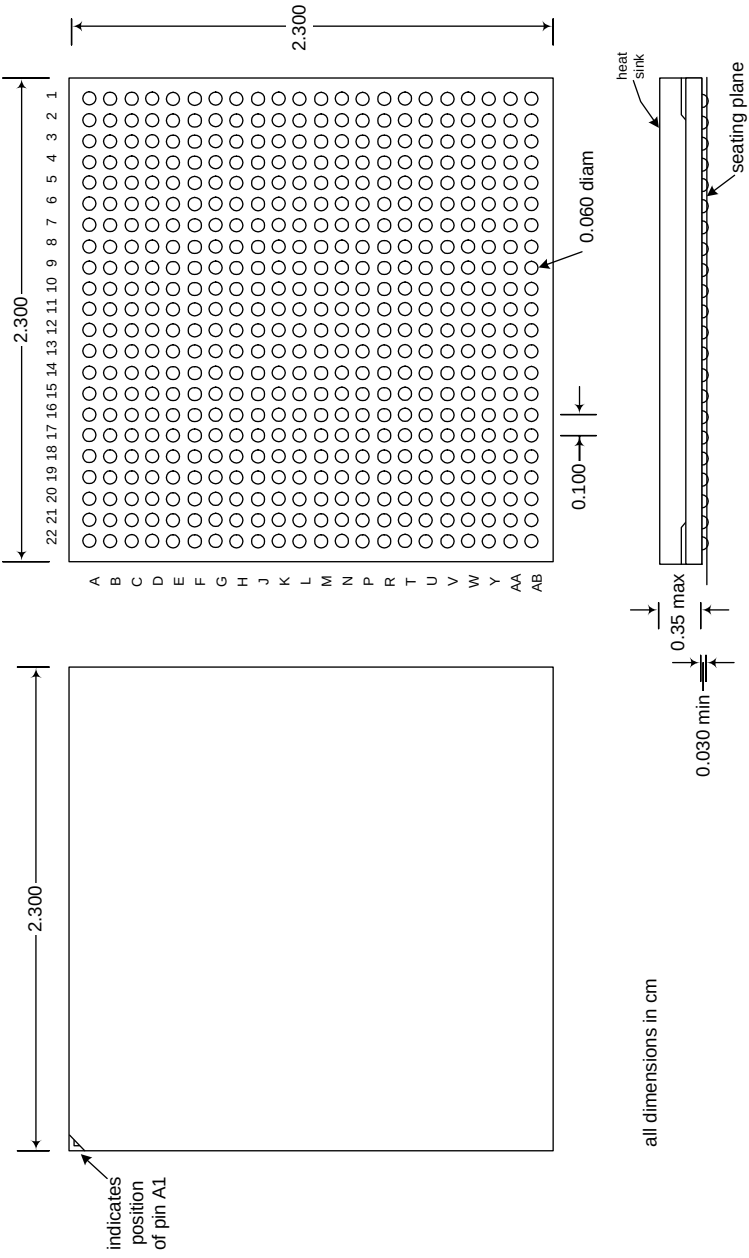


FIGURE 18.4
Ball grid array (BGA) package with 484 pins and an integrated heat sink.

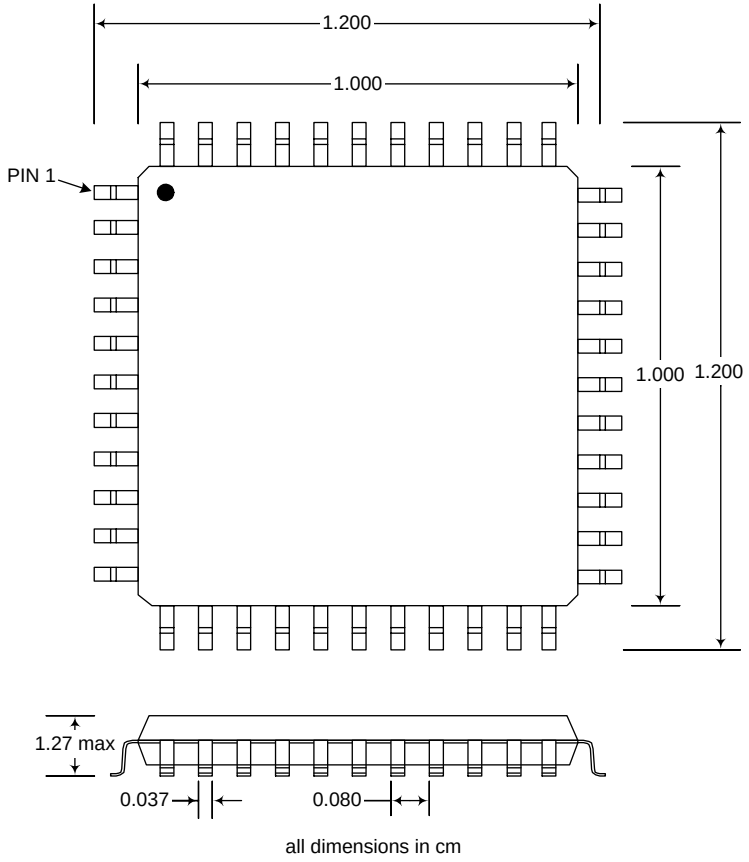


FIGURE 18.5
Plastic quad flat pack (QFP) package with 44 pins.

one die in a single package. Some modules use die arranged in a single plane whereas others stack the chips vertically to reduce the package footprint significantly. An example of the latter approach is the memory cube, in which dynamic random access memories (DRAMs) are stacked vertically. In either case, the board area consumed is significantly less than if the die were packaged individually.

Multichip modules with a number of material technologies have been developed; each has its own cost–performance trade-off. MCM-C technology utilizes ceramic-based substrates. The ceramic layers are laminated together with many levels (~50) of metallization. MCM-D technology uses layers of deposited metal and insulator layers to achieve thinner layers and superior lead pitch, resulting in highest performance but also highest cost. MCM-L technology utilizes laminated organic layers such as polyamide for reduced cost.

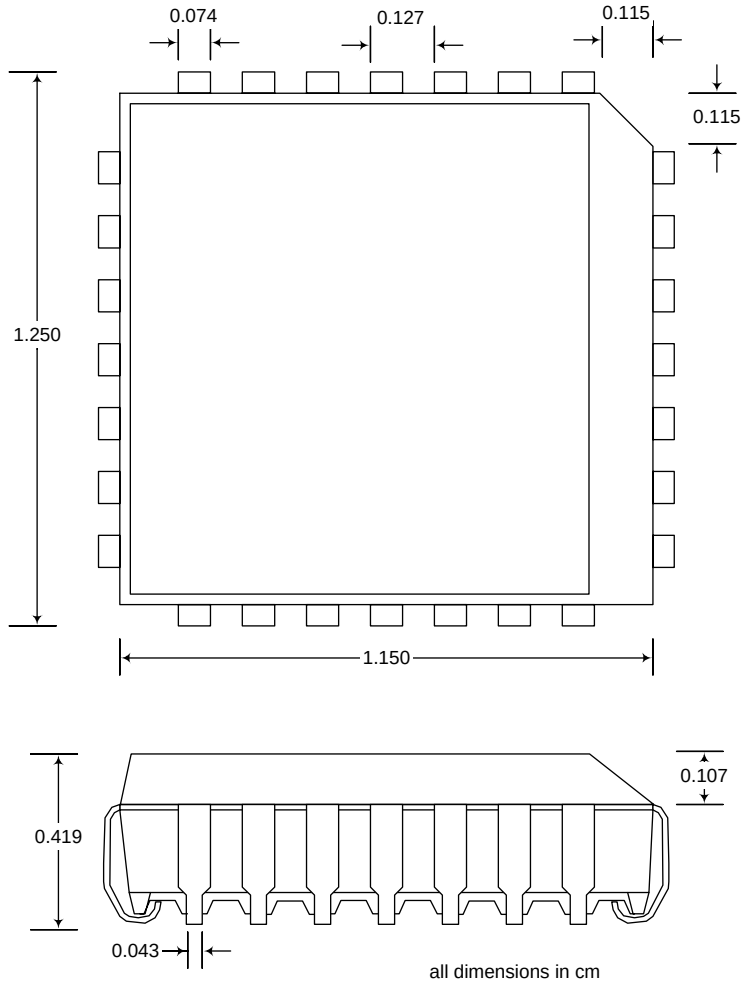


FIGURE 18.6
Plastic J-lead chip carrier (PLCC) with 28 pins.

18.2.6 Trends in Package Types

Market demands for higher-density integrated circuits with increased functionality, higher off-chip data rates, and higher power densities have reduced the use of through-hole technologies such as dual in-line packages and pin-grid arrays. On the other hand, the overall market for integrated circuits has grown explosively. This has led to the increased use of conventional surface mount technology, ball grid arrays, and bare die (chip on board). These trends can be seen in Figure 18.7.

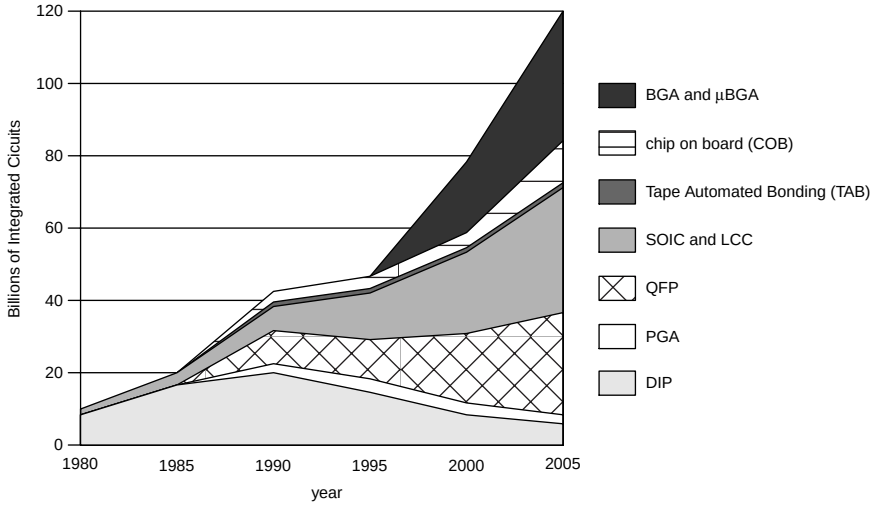


FIGURE 18.7
Integrated circuit package usage by type. (www.altera.com [Altera Corporation].)

18.3 General Considerations

There are several general requirements for any integrated circuit package; such a package must be able to:

- Provide an adequate number of electrical connections to the outside world (usually called pins) without imposing long signal delays
- Conduct heat sufficiently from the operating circuit
- Withstand elevated temperatures imposed by the circuit operation
- Withstand the thermal cycling associated with normal circuit operation without imposing mechanical failure due to thermal stresses
- Protect the integrated circuit from the chemical environment, especially moisture and ionic contaminants
- Protect the circuit from mechanical vibration, mechanical shock, and stresses
- Be easily handled, tested, and assembled into systems

These requirements can be broadly categorized as electrical, thermal, chemical, and mechanical.

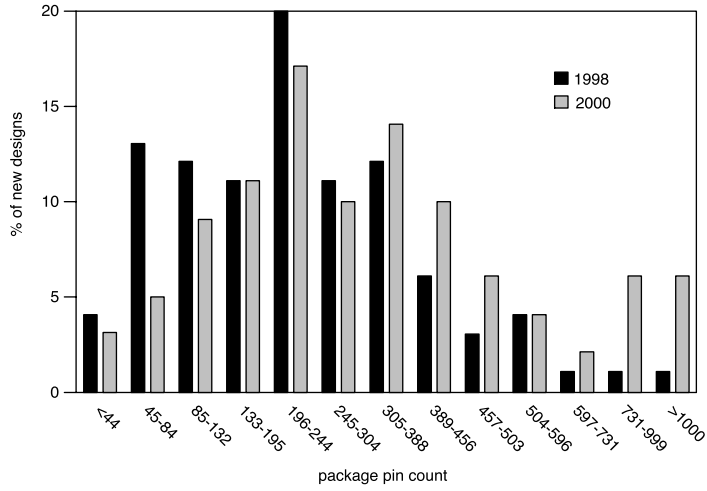


FIGURE 18.8
Percentage of application-specific integrated circuit (ASIC) starts vs. pin count (www.altera.com [Altera Corporation].)

18.3.1 Electrical Considerations

An integrated circuit package must provide the required electrical connections without imposing undue signal delays due to parasitic inductances, capacitances, or resistances in the leads.

VLSI circuits often require pin counts in the hundreds, a requirement that is increasing steadily. This is illustrated for the case of application-specific integrated circuits (ASICs) in Figure 18.8.[5] Therefore, areal pin density is a commonly used figure of merit for packages. Packages with leads arranged in straight rows along two sides (such as DIPs) have low pin density; packages utilizing square grids of pins or metal bumps have much higher pin densities (>30 pins/cm^2).

When modeling the electrical behavior of package pins, the common practice is to choose between a transmission line model and a lumped element model. Although these two models represent limiting cases, they greatly simplify the analysis and often provide reasonable accuracy. The choice between the two models is made based on a comparison between the propagation delay of the circuit and the time of flight for the electrical signal. The time of flight is given by

$$t_{flight} = \frac{l}{c_0 / \sqrt{\varepsilon_r \mu_r}}, \tag{18.1}$$

where l is the electrical path length, c_0 is the speed of light in free space, ε_r is the relative permittivity for the medium, and μ_r is the relative permeability

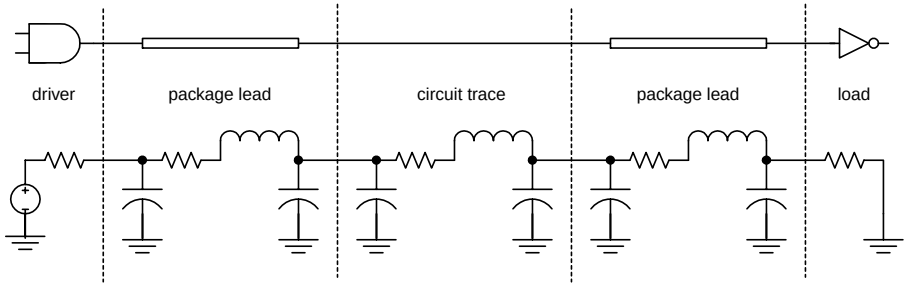


FIGURE 18.9
Lumped element model for a packaged circuit driving another packaged circuit.

TABLE 18.1

Typical Parasitics and Signal Delays Associated with Two Different Package Approaches

	W/B[a] w/PGA[b]	F/C[c] w/BGA[d]
Inductance	10 nH	1.5 nH
Capacitance	12 pF	4 pF
Resistance	20 Ω	2 Ω
Lead signal delay	700 ps	100 ps

[a] W/B = wafer bonded.
[b] PGA = pin grid array.
[c] F/C = flip chip.
[d] BGA = ball grid array.

Source: Blackwell, G.R., *The Electronic Packaging Handbook*, CRC Press, Boca Raton, FL, in cooperation with IEEE Press, 2000.

for the medium. If the circuit propagation delay is less than the time of flight, then the transmission model should be used. Otherwise, a lumped element model is applicable. In practice, the lumped element model can often be used for traces on circuit boards, whereas the transmission line model must be employed for network connections.

Figure 18.9 illustrates the use of a lumped element model for a case in which a packaged circuit drives the input to another packaged circuit. Table 18.1 provides some typical values of the parasitics associated with package leads.[1] These numbers, though specific to two particular packaging approaches, demonstrate the importance of minimizing the package parasitics for high-performance applications.

For insulating materials used in packages, it is desirable to have low values of the dielectric constant and the loss tangent (Table 18.2). The power dissipation and development of heat in the insulator are directly proportional to the loss tangent (also referred to as the dissipation factor). In order to reduce the parasitic capacitances associated with the integrated circuit package, it is desirable to use materials with lower dielectric constants. Quartz is superior to the other ceramics in this regard. Epoxy resin, used in plastic packages,

TABLE 18.2

Relative Permittivities (Dielectric Constants) and Loss Tangents of Insulating Materials Used in Digital Integrated Circuit Packages

Material	Dielectric Constant ε_r @ 1 MHz	Loss Tangent ($\times 10^4$) @ 25°C, 1 MHz
Polyamide	3.4–4.0	0.0025–0.01
Epoxy resin	3.5–4.0	300
Quartz	3.5–4.0	2
Si_3N_4	6–10	—
Beryllia	6.7–8.9	4–7
AlN	8.5–10	5–10

Source: Pecht, M.G., Agarwal, R., McCluskey, P., Dishongh, T., Javadpour, S., and Mahajan, R., *Electronic Packaging Materials and Their Properties*, CRC Press, Boca Raton, FL, 1999.

TABLE 18.3

Electrical Resistivities of Conductors Commonly Employed in Integrated Circuit Packages

Metal	ρ ($\mu\Omega$ cm)
Copper (Cu)	1.7
Gold (Au)	2.2
Aluminum (Al)	2.65

also has a similar dielectric constant but is relatively lossy. These properties of packaging insulators are summarized in Table 18.2.

For conducting materials, smaller values of the electrical resistivity are desirable because they give rise to smaller parasitic resistances. As can be seen in Table 18.3, copper is superior in this regard and finds use in substrate conductors. Aluminum is used almost exclusively for bonding pads, whereas gold and aluminum have been used for wire bonds.

Example 18.1

Suppose a packaged digital circuit drives another package's circuit. There is a 4-cm trace between them on an FR-4 circuit board. What is the minimum rise time for which a lumped element model can be used?

Solution. For FR-4, the relative permittivity is 4. The propagation time of flight is

$$t_{flight} = \frac{4 \text{ cm}}{3 \times 10^{10} \text{ cms}^{-1} / \sqrt{4}} = 67 \text{ ps}.$$

Therefore, the critical rise time is 134 ps.

18.3.2 Thermal Considerations

The important thermal considerations are heat dissipation[36–38] and thermal expansion. Efficient heat removal is necessary in order to minimize the junction temperatures of the operating circuits, to avoid malfunction or irreversible failure. Junction leakage currents increase exponentially with temperature. Most integrated circuit failure mechanisms are also thermally activated, so circuit lifetimes decrease strongly with operating temperature. Thermal expansion must be considered because the integrated circuit package utilizes many disparate materials with very different thermal expansion coefficients. Thermal cycling of the packaged circuit therefore gives rise to thermal stresses; in turn these may result in failure during circuit board assembly or normal operation of the circuit.

Conductive heat flow in a solid is governed by the Fourier equation:

$$q = -k\nabla T \,, \tag{18.2}$$

where q is the heat flow in W/cm^2, k is the thermal conductivity of the solid in $Wcm^{-1}K^{-1}$, and ∇T is the three-dimensional temperature gradient in K/cm. In a one-dimensional case, the heat flow can be described by an equation analogous to Ohm's law using the thermal resistance. For a layer of a solid with a cross-sectional area of A, a thickness of l, and a thermal conductivity of k, the thermal resistance is given by

$$\theta = \frac{l}{KA} \,. \tag{18.3}$$

The one-dimensional heat flow is given by

$$Q = \frac{\Delta T}{\theta} \,, \tag{18.4}$$

where Q is the heat flow in W (analogous to electrical current), ΔT is the temperature difference in K (analogous to potential difference) and θ is the thermal resistance in W/K (analogous to electrical resistance).

For a dissipating integrated circuit, the junction temperatures can be calculated from

$$T_j = T_a + P_d\theta_{ja} \,, \tag{18.5}$$

where T_j is the junction temperature, T_a is the ambient temperature, P_d is the power dissipated by the chip, and θ_{ja} is the junction-to-ambient thermal resistance. Often this thermal resistance comprises a number of series components. In such a case,

$$\theta_{ja} = \theta_1 + \theta_2 + \theta_3 + \ldots \tag{18.6}$$

TABLE 18.4

Thermal Conductivities of Materials Commonly Used in Integrated Circuit Packages

Material	K (W/mK)
Semiconductors	
Silicon carbide (SiC)	90–260
Silicon (Si)	150
Gallium arsenide (GaAs)	50
Substrate materials	
Diamond (C)	2000
Beryllia (BeO)	260–300
Aluminum nitride (AlN)	100–270
Alumina 96% (Al_2O_3)	30
Metals	
Silver (Ag)	428
Copper (Cu)	397
Gold (Au)	317
Aluminum (Al)	230
Nickel (Ni)	88

Source: Pecht, M.G., Agarwal, R., McCluskey, P., Dishongh, T., Javadpour, S., and Mahajan, R., *Electronic Packaging Materials and Their Properties*, CRC Press, Boca Raton, FL, 1999.

More complicated situations are also encountered; however, the thermal resistances combine in the same manner as electrical resistances.

The thermal conductivities of materials commonly used in integrated packages are tabulated in Table 18.4. Silicon has three times the thermal conductivity of GaAs; therefore, GaAs integrated circuits often require mechanical thinning to achieve efficient heat transfer. Diamond exhibits superior thermal conductivity compared to other substrate materials but is only used in high-power applications because of its expense.

In silicon circuits, the maximum allowable junction temperature for an operating circuit is 125°C. This places an upper limit on the package thermal resistance in any given application; however, lower junction temperatures enhance the die reliability.

A number of package design strategies have been used to reduce thermal resistance. For example, electrical pins provide important pathways for heat conduction away from the circuit. Therefore, the arrangement of many pins in a grid covering the bottom of the package provides superior heat removal compared to using pins at the package periphery. Also, because the circuitry resides in the top 1% of the wafer thickness, it is beneficial to mount the

package face down. This approach, called flip-chip technology, places the dissipating transistors in closer contact with the substrate. In some VLSI applications, such as microprocessors, the dissipation is such that a metal heat sink must be built into the package. Forced convection, either air or liquid, may also be used.

18.3.3 Chemical Considerations

An integrated circuit package must protect the circuit from its chemical environment during storage and operation. Also, the many materials used in its manufacture must be chemically compatible.

In most applications, water vapor is the most important environmental concern. Many packaging materials are hygroscopic; thus, parts stored for any duration of time will soak up appreciable amounts of water from the air. If these parts are not baked out adequately prior to assembly, the sudden temperature rise associated with soldering will cause package failure (*popcorning*)[39–43] due to rapid vaporization of the water. During operation of assembled systems, contamination by water and ionic contaminants will cause gradual circuit degradation despite the use of encapsulants (cover layers) over the circuits. Examples of encapsulants include phosphosilicate glass, polyamide, silicon nitride (deposited *during fabrication*), or silicone (deposited *after fabrication*). A costly but effective means for eliminating these problems is to use a hermetically sealed package.

Hermetically sealed packages are constructed using metal, ceramic, or metal–ceramic enclosures with glass seals. These enclosures block the migration of water and other contaminants into the package and have been commonly used for aerospace and military applications.

One popular hermetic package uses Kovar (a metallic alloy of 54% Fe, 29% Ni, and 17% Co). Here, the Kovar package and lid are hermetically sealed using a glass frit. The thermal expansion coefficient of Kovar (5.1 to 5.9 ppm/K) closely matches thermal expansion coefficients of commonly used sealing glasses (5.25 to 6.96 ppm/K), therefore minimizing thermal stresses associated with the seal. Because the thermal conductivity of Kovar is relatively poor (15.5 to 17 W/mK), copper alloys are used for high-power hermetic packages.

18.3.4 Mechanical Considerations

Generally speaking, mechanical failure mechanisms in an integrated circuit package may be classified as an instantaneous mechanical overload or as progressive in nature.[44] Instantaneous overloading problems include ductile deformation and brittle fracture; progressive failure mechanisms include fatigue crack growth and creep deformation.

TABLE 18.5

Yield Stresses of Metals Commonly Used in Integrated Circuit Packages

Metal	σ_y (Mpa)
Nickel (Ni)	70
Copper (Cu)	60
Aluminum (Al)	40
Gold (Au)	40
Lead (Pb)	11
63% Lead–37% tin solder	~10

Ductile overload occurs in metals such as aluminum, copper, gold, and solder when the critical stress is exceeded. Here, the stress σ is a simple function of the applied force F and the cross sectional area A for the metal element:

$$\sigma = \frac{F}{A}. \tag{18.7}$$

If the applied stress exceeds the yield stress, σ_y, of the metal, permanent deformation will occur, possibly resulting in a broken electrical connection. In practice, packaging engineers must ensure that the yield stress is never exceeded. Table 18.5 gives the yield stresses of metals commonly used in integrated packages; it is noteworthy that eutectic lead-tin solder is especially poor in this regard.

Brittle materials such as ceramic substrates may fail by fracture at a point at which there is an existing flaw in the material. The stress associated with brittle fracture is given by

$$\sigma = \frac{YK_{lc}}{\sqrt{a}}, \tag{18.8}$$

where K_{lc} is the fracture toughness of the material, a is the size of the relevant flaw, and Y is a constant of proportionality. The values of fracture toughness for materials commonly used in integrated packages are summarized in Table 18.6, which shows that beryllia, alumina, and silicon carbide are superior in this regard.

Time-dependent, progressive failure mechanisms include fatigue crack growth and creep deformation. Fatigue crack growth occurs by repeated stress cycles of the type present during the normal thermal cycling of an integrated circuit. Creep occurs under a constant high-stress condition at elevated temperature. Here, plastic deformation gradually increases over a long period of time, thus giving rise to failure. Of these, fatigue crack growth is the more common phenomenon because the normal power-up, power-down cycling of integrated circuits gives rise to cyclic thermal strains. In the absence of cracks, the fatigue of such materials may be described by *Basquin's*

TABLE 18.6

Fracture Toughnesses of Metals Commonly Used in Integrated Circuit Packages

Material	K_{Ic} (Mpa $m^{1/2}$)
Silicon carbide (SiC)	3–3.5
Alumina (Al_2O_3)	3
Silica glass (SiO_2)	0.5
Fused quartz (SiO_2)	0.5
Beryllia (BeO)	3.7
Conductive epoxy	0.3–0.5

law. This empirical relationship states that the lifetime of a material subjected to a repeated cycle of stress (*below* the yield point) is given by

$$L_o = B(\Delta\sigma)^{-\theta_b}, \tag{18.9}$$

where $\Delta\sigma$ is the peak-to-peak amplitude of the time-varying stress and B and θ_b are material parameters; typically, $8 < \theta_b < 15$. Usually, the periodic stress results from thermal cycles. As a consequence, fatigue lifetimes are often stated in terms of the temperature cycling (which can be more directly measured) rather than the period stress. Fatigue is an important failure mechanism for solder bumps used in flip-chip technology and for wire bonds in plastic packages.

18.4 Packaging Processes and Materials

The process of packaging an integrated circuit involves many steps and very different materials, each with properties to serve its specialized purpose; wire bonding or the flip-chip approach may be used. Wire bonding involves mounting the die face up and running wires from the die to the pins. The flip-chip approach places the die face down so that electrical connections between the die and package are made by solder bumps. The materials used in the packaging process include metals, ceramics, glasses, and organics. Metals are used for pins, wires, solder bumps, and package enclosures; ceramics as substrates and package enclosures; glasses to seal hermetic enclosures made of ceramic or metal; and organics for encapsulants, molded plastic packages and form adhesives.

18.4.1 Wire-Bond Process

The wire-bonding approach is commonly used in conjunction with plastic and ceramic packages. By this process, electrical connections between the

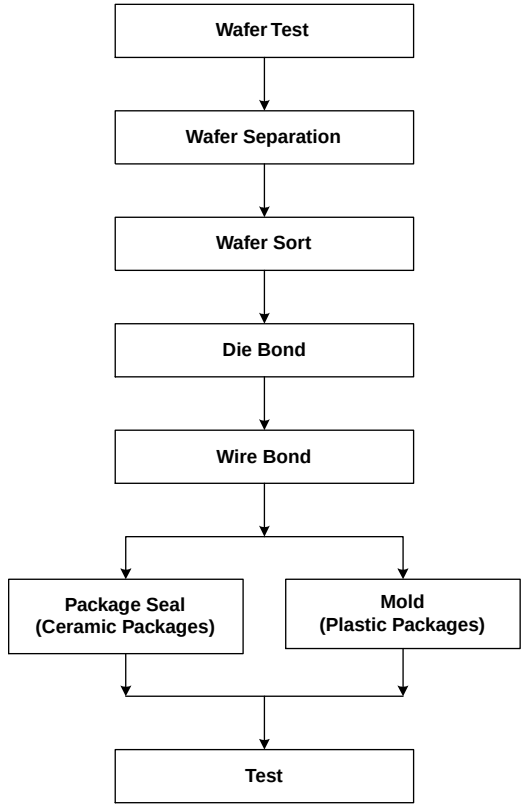


FIGURE 18.10
Process flow for a wire bond packaging process.

die and the pins are made using fine gold or aluminum wires. The process flow for a wire bond process is outlined in Figure 18.10. Following wafer fabrication, the individual circuits are subjected to an electrical test on the wafer (*wafer test*) using a *wafer probe*. Failed circuits are marked with a dot of ink so that they may be discarded. Next, the die are separated by cleaving* or sawing with a diamond saw (*wafer separation*); bad die are discarded at the *wafer sort* step. Thus, only the known good die (KGD) are packaged, resulting in considerable cost savings.

Die bonding involves attaching known good die to a ceramic substrate or a metal lead frame. Ceramic substrates are commonly alumina–silica mixtures (90 to 99% Al_2O_3, balance SiO_2) or beryllia (BeO); however, many other materials are available. Metal lead frames are made from a copper alloy or Kovar (a metallic alloy of 54% Fe, 29% Ni, and 17% Co). Electrical connections are made from the die to the package leads by wire bonding.

* Cleavage is the separation of the crystal along natural crystal planes, called "cleavage planes." For example, silicon crystals cleave on [111] planes. For the case of a silicon [001] wafer, cleavage on these planes' results is rectangular die, with edges oriented by 54.7° to the top surface.

The wire-bonding process has several variations. One method in common use is the thermosonic ball-wedge technique[45] illustrated in Figure 18.11. In this process a fine gold wire is drawn through a tungsten carbide capillary. A round ball is produced on the end of the wire by a hydrogen microtorch or by capacitive discharge. Either case results in localized melting of the gold to form a ball. Next, this ball is welded to the aluminum bonding pad using a combination of downward pressure, heat (~150°C) and ultrasonic vibration (~50 kHz). The ultrasonic vibration serves to break up the tough native oxide layer on the aluminum pad. The combination of heat and pressure promotes localized melting and therefore welding of the gold to the aluminum.

Following this, the tool tip is pulled to a position over the metal lead, drawing a length of gold wire from the capillary. Then, the tool tip is pressed down on the metal lead, with heat and ultrasonic vibration. When the tool is drawn away at a shallow angle, the wire breaks to form a wedge bond. At this point the tip is ready to make the next wire bond. This process is repeated until all connections have been made. Following the wire bonding process, the integrated circuit is enclosed by a transfer molding process (plastic packages) or a package seal process (ceramic packages). The transfer molding process involves placing a measured quantity of the molding compound in a metal mold. The thermosetting molding compound melts and conforms to the shape of the package mold under the applied pressure (~6 Mpa) and heat (~175°C).

Molding compounds in common use include novolac epoxies, silicone, and epoxy silicone. Usually these molding compounds are loaded (~70% by weight) with a filler such as SiO_2 or Al_2O_3, resulting in a material with improved thermal characteristics (expansion coefficient and thermal conductivity). The molding process is particularly hard on wire bonds. This is because the molding compound surrounds the bond wires prior to hardening, which is accompanied by the introduction of mechanical stress. For this reason, preformed plastic packages are sometimes used.

The package sealing process involves the bonding of a metal or ceramic lid on the ceramic substrate using an intermediate glass layer. Glasses used for this purpose are $PbO/ZnO/B_2O_3$ mixtures of various compositions. These glasses flow at 400°C, forming a hermetic seal. However, the relatively high temperature involved necessitates the use of aluminum bond wires to avoid gold–aluminum reactions. After the molding or package sealing process, the packaged circuits are subjected to electrical tests so that defective devices can be identified and discarded. Burn-in and thermal cycle testing are also used so that short-lifespan units can be rejected.

18.4.2 Flip-Chip Process

The starting die for a flip-chip process must be fabricated with solder bumps to facilitate electrical connection to the package. Typically, these solder bumps are made using a Pb–Sn eutectic or a Pb–In alloy. Figure 18.12 illustrates the implementation of a Pb–Sn solder bump over an aluminum pad.

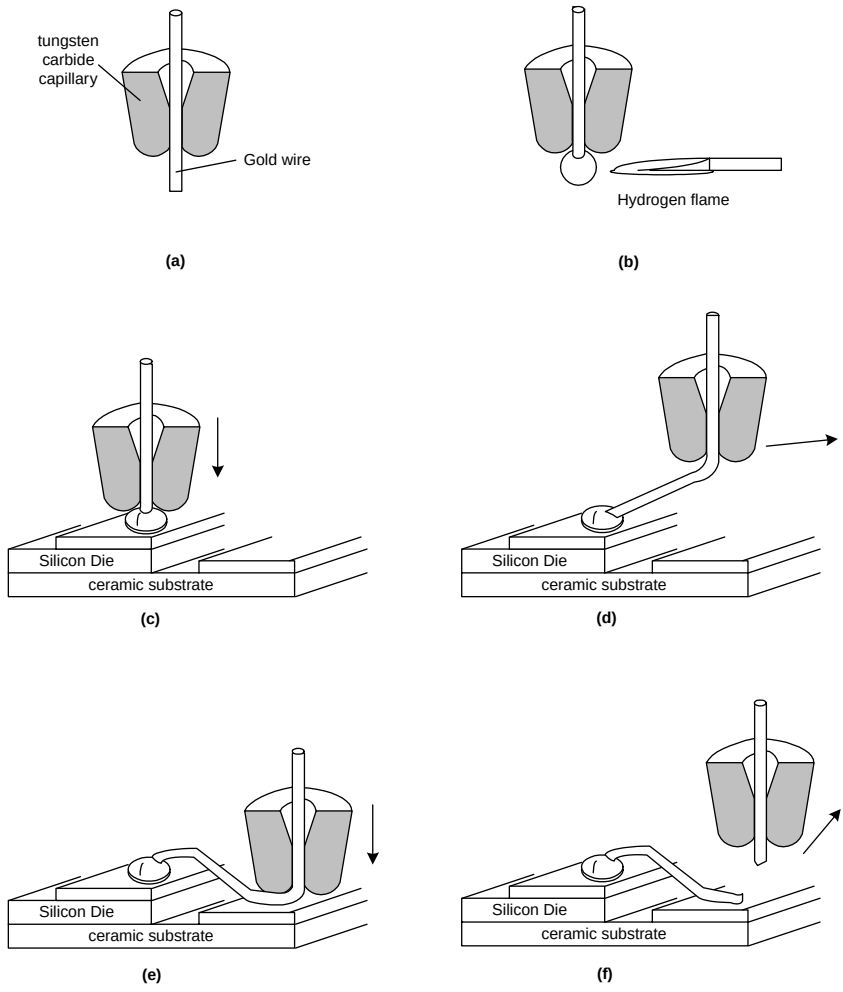


FIGURE 18.11
Wire bonding using the thermosonic ball-wedge approach. (a) A gold wire, 10 to 50 μm in diameter, is drawn through a tungsten carbide capillary. (b) A hydrogen microtorch or a capacitive discharge is used to form a gold ball on the end of the wire. (c) The gold ball is bonded to an aluminum pad on the heated (~150°C) silicon die using a vertically applied force and ultrasonic vibration (~50 kHz). (e) The capillary tip is pulled over to the metal pad on the heated substrate; bonding is achieved using a vertical force with ultrasonic vibration. (f) The tip is pulled away, breaking the gold wire, which is ready for the next wire bond. (Adapted from Ghandhi, S.K., *The Theory and Practice of Microelectronics*, Robert E. Krieger, Malabar, FL, 1968.)

The process flow for the flip-chip approach is outlined in Figure 18.13. Bumped wafers undergo wafer test, wafer separation, and wafer sort as described before. Then the bumped die is flipped over, face down, on the substrate. The solder bumps mate to metal lands on the package. Solder reflow is conducted at an elevated temperature (230°C for Pb–Sn eutectic)

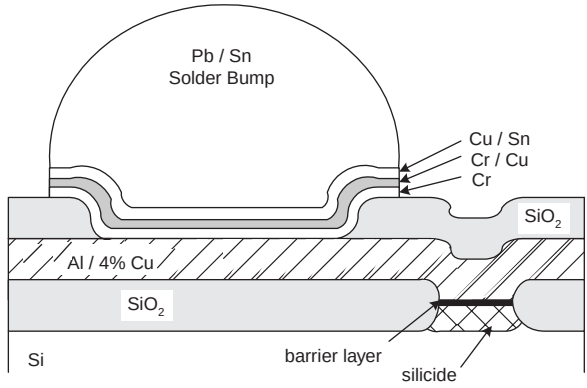


FIGURE 18.12
A solder bump on a silicon wafer.

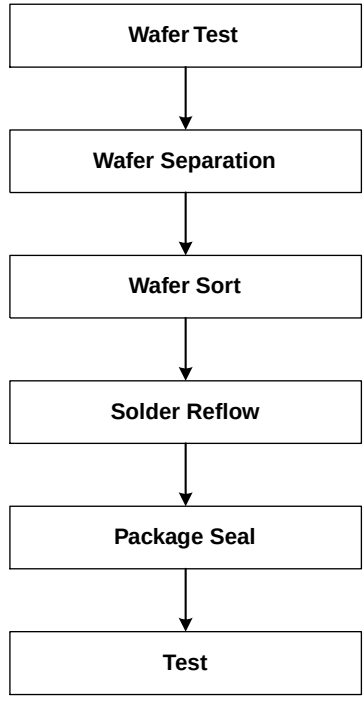


FIGURE 18.13
Process flow for a flip-chip packaging process.

that forms an excellent electrical and mechanical connection between the flip chip and the package; the surface tension of the molten solder insures proper alignment between them. Following reflow, the package is sealed and, finally, the packaged circuit is tested.

TABLE 18.7

Thermal Fatigue Lifetimes for Solder Bump Alloys[a] Normalized to Lifetime for Eutectic Tin–Lead Solder[b] with No Underfill

Bump Alloy	T (reflow) (°C)	Normalized Life (no underfill)	Normalized Life (epoxy underfill)
63Sn/37Pb	230	1.0	15
50In/50Pb	260	2–3	>30
37In/63Pb	290	2–3	>30
3.5Ag/96.5Sn	260	0.5	11
5Sb/95Sn	280	0.3	11
Sn/Pb/Cd/In	230	1.0	13
Sn/Ag/Cu/Sb	260	1.0	13

[a] –40 to +125°C cycles.
[b] 63% Sn/37% Pb.

Elimination of the bonding wires in flip-chip packages allows bonding pads that cover the entire chip area, rather than just the periphery, to be used. In addition to greater pin density and improved heat removal, this avoids the signal delays associated with inductance and resistance of the bonding wires. A problem encountered in flip-chip packages is the thermal fatigue of the solder bump connections. Using epoxy underfill with the solder bumps greatly enhances the fatigue lifetimes of solder bump connections, as is evident from the results compiled in Table 18.7 for common solder bump alloys.

18.5 Summary

Once digital integrated circuits have been designed and fabricated on a wafer, the wafer is cut into rectangular die that are tested and packaged for assembly in systems. Packaging requirements for VLSI circuits are rather stringent, requiring large numbers (~10^3) of electrical connections, capability of high input and output data rates (~10^9 bits/s), and efficient removal of large quantities of heat (~10^2 W). Moreover, these packages must be compact, lightweight, inexpensive, and reliable. The five basic types of integrated circuit packages are through-hole packages, surface mount packages, chip-scale packages, bare die, and module assemblies.

Through-hole technology (THT) packages have metal pins that may be inserted through holes drilled in the circuit board for soldering. Surface mount technology (SMT) packages utilize metal leads that can be soldered to a single surface of the printed circuit board. They are much smaller and more lightweight than through-hole packages for a given number of electrical

connections. In addition, they are more resistant to mechanical shock compared to through-hole parts. Chip-scale packages represent the most compact packaging scheme apart from the use of bare die. Typically, the package dimensions are only 20% greater than the die dimensions. However, chip-scale packages offer advantages in handling and testability compared to bare die. Usually, these packages are attached to circuit boards via an array of metal bumps. This technology provides a high pin density and is mechanically robust. Bare or *unpackaged* parts offer the minimum size and weight and also eliminate the RC time delays associated with the package leads. Module assemblies combine bare die, or occasionally packaged die, in a module. Some modules use stacked die to achieve the minimum connection lengths and the highest efficiency in circuit board utilization.

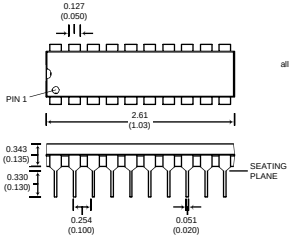
The process of packaging an integrated circuit involves wire bonding or the flip-chip approach. Wire bonding involves mounting the die face up and running wires from the die to the pins. The flip-chip approach places the die face down so that electrical connections between the die and package are made by solder bumps. The materials used in the packaging process include metals, ceramics, glasses, and organics. Metals are used for pins, wires, solder bumps, and package enclosures; ceramics are used as substrates and package enclosures. Glasses are used to seal hermetic enclosures made of ceramic or metal. Organics are used for encapsulants, molded plastic packages and form adhesives.

INTEGRATED CIRCUIT PACKAGES QUICK REFERENCE

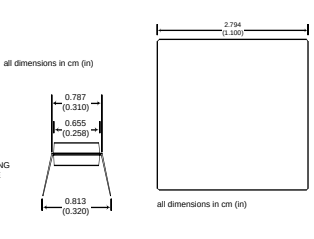
Integrated Circuit Packages

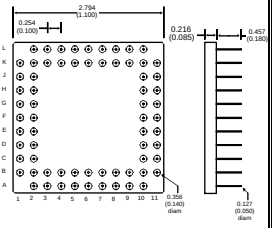
There are five approaches to packaging integrated circuits: through-hole packages, surface-mount packages, chip scale packages, bare die, and multichip modules. Through-hole packages include DIPs and PGAs. Surface mount packages include SOICs, QFPs, PLCCs, and BGAs. Chip scale packages include μBGAs. ICs may be packaged inexpensively by wire bonding. Higher performance packages use flip-chip technology. The design of IC packages and the systems incorporating them requires the consideration of electrical, thermal, mechanical, and chemical considerations.

Through-Hole Packages



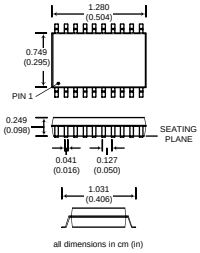
20 pin PDIP



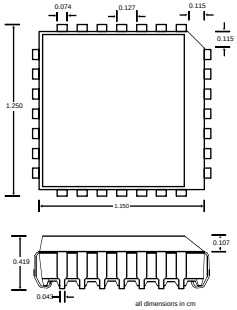


68 pin PGA

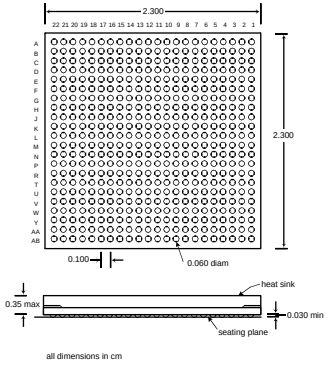
Surface Mount Packages



20 pin SOIC



28 pin PLCC



484 pin BGA

Electrical Considerations

The lumped circuit model is applicable if $t_{rise} > 2t_{flight}$.

The transmission line model is applicable if $t_{rise} < 2t_{flight}$.

$$t_{flight} = \frac{l}{c_0 / \sqrt{\varepsilon_r \mu_r}},$$

Thermal Considerations

$$T_j = T_a + P_d \theta_{ja}$$

f = 10^{-15} p = 10^{-12} n = 10^{-9} μ = 10^{-6} m = 10^{-3} k = 10^{3} M = 10^{6} G = 10^{9}

Problems

P18.1. Calculate and compare the pin densities (in pins/cm^2) for a 20-pin PDIP, a 20-pin SOIC, and a 655-pin BGA.

P18.2. Consult the MOSIS Web site (www.mosis.org) to determine the minimum dimensions and spacings for bonding pads. Using this information, determine the practical maximum number of wire bonds that can be made to a 2 cm $\times$ 2 cm die.

P18.3. The occurrence of popcorning in a 20-pin plastic DIP corresponds to the evolution of 10^{-4} cm^3 of water vapor. What is the corresponding mass of absorbed water?

P18.4. The output pins of a packaged CMOS microprocessor (rise time ~1 ns) drives other circuits on an FR-4 motherboard. The electrical paths vary from 2 to 20 cm. Is the lumped element model applicable to the analysis of the signal delays?

P18.5. A packaged emitter-coupled logic (ECL) gate with a rise time of 100 ps drives a 20-m network cable with a dielectric constant of 3. Which is applicable to the analysis: a lumped element model or a transmission line model?

P18.6. When mounted on an FR-4 printed circuit board, a 44-pin PLCC package has a junction to ambient thermal resistance of 65°C/W. What is the maximum allowable dissipation for a silicon circuit housed in this circuit if the ambient temperature is 25°C?

P18.7. When mounted on an FR-4 printed circuit board, a flip chip in a 680-pin ceramic PGA package has a junction to ambient thermal resistance of 8.2°C/W. What is the maximum allowable dissipation for a silicon circuit housed in this circuit if the equipment cabinet temperature is 45°C?

P18.8. A silicon microprocessor with a maximum dissipation of 95 W is mounted in a 480-pin BGA package with an integrated heat sink. The junction to ambient thermal resistance is 3.5°C/W. Is forced convection necessary?

References

1. Blackwell, G.R., *The Electronic Packaging Handbook*, CRC Press in cooperation with IEEE Press, Boca Raton, FL, 2000.
2. www.jedec.org (Joint Electron Device Engineering Council, or JEDEC).
3. www.ieee.org (Institute of Electrical and Electronics Engineers).
4. www.intel.com (Intel Corporation).
5. www.altera.com (Altera Corporation).

6. Howell, J.R., Reliability study of plastic encapsulated copper lead frame epoxy die attach packaging system, *Proc. Int. Reliability Phys. Symp.*, 104, 1981.
7. Levinthal, D.S., Semiconductor packaging trends, *Semicond. Int.*, 33, April 1979.
8. Peng Yeoh, H.P., Lii, M.-L., Sankman, B., and Azimi, H., Flip chip pin grid array (FC-PGA) packaging technology, *Proc. 3rd Electron. Packag. Technol. Conf.*, 33, 2000.
9. Miwa, T., Otsuka, K., Shirai, Y., Matsunaga, T., and Tsuboi, T., High reliability and low cost in plastic PGA package with high performance, *Proc. 41st Electron. Components Technol. Conf.*, 183, 1991.
10. Knausenberger, W. and Teneketges, N., High pinout IC packaging and the density advantage of surface mounting components, *IEEE Trans. Hybrids, Manuf. Technol.*, 6, 298, 1983.
11. Mattei, C. and Agrawal, A.P., Electrical characterization of BGA packages, *Proc. 47th Electron. Components Technol. Conf.*, 1087, 1997.
12. Lin, P. and McShane, M., Approaches to high pin count and high power surface mount packages, *Proc. 1991 Int. Symp. VLSI Technol., Syst., Appl.*, 141, 1991.
13. Freyman, B. and Marrs, R., Ball grid array (BGA): the new standard for high I/O surface mount packages, *Proc. 1993 Jpn. Int. Electron. Manuf. Technol. Symp.*, 41, 1993.
14. Rao, S.T., Ball grid array assembly issues in manufacturing, *Proc. 16TH IEEE/CPMT Int. Electron. Manuf. Technol. Symp.*, 347, 1994.
15. Thompson, P., Chip-scale packaging, *IEEE Spectrum*, 34, 36, Aug. 1997.
16. Okuno, A., Fujita, N., and Ishikana, Y., Low cost and high reliability extremity CSP packaging technology, *Proc. 49th Electron. Components Technol. Conf.*, 1201, 1999.
17. Elenius, P., The ultra CSP™ wafer scale package, *Proc. 2nd Electron. Packag. Technol. Conf.*, 83, 1998.
18. Arnold, R., Chip scale package versus direct chip attach (CSP vs. DCA), *Proc. 50th Electron. Components Technol. Conf.*, 822, 2000.
19. Bauer, C.E., Micro/chip scale packages and the semiconductor industry road map, *Proc. 2nd IEMT/IMC Symp.*, 302, 1998.
20. Intel flash memory chip scale package user guide, Intel Corporation application note, www.intel.com, 1999.
21. Rochat, G., COB and COC for low cost and high density package, *Proc. 17th IEEE/CPMT Int. Electron. Manuf. Technol. Symp.*, 109, 1995.
22. Ganasan, J.R., Chip on chip (COC) and chip on board (COB) assembly on flex rigid printed circuit assemblies, *Proc. 49th Electron. Components Technol. Conf.*, 174, 1999.
23. Santeusanio, D., Bare die tape and reel for high volume manufacturing, *Proc. Electro 1999*, 87, 1999.
24. Fillion, R., Burdick, B., Shaddock, D., and Piacente, P., Chip scale packaging using chip-on-flex technology, *Proc. 47th Electron. Components Technol. Conf.*, 638, 1997.
25. O'Malley, G., Giesler, J., and Machuga, S., The importance of material selection for flip-chip on-board assemblies, *Proc. 44th Electron. Components Technol. Conf.*, 387, 1994.
26. Understanding the quality and reliability requirements for bare die applications, Micron Technology, Inc. technical note, www.micron.com, 2002.
27. Charles, H.K., Packaging with multichip modules, *Proc. 13th IEEE/CHMT Electron. Manuf. Technol. Symp.*, 206, 1992.

28. Vasquez, B. and Tippins, F., Multichip modules: packaging solutions for size performance integration, *Int. Integrated Reliability Workshop Final Rep.*, 215, 1993.
29. Crowley, R.T. and Vardaman, E.J., 3-D multichip packaging for memory modules, *Proc. 1994 Int. Conf. Multichip Modules*, 474, 1994.
30. Simsek, A. and Reichl, H., Evaluation and optimization of MCM-BGA packages, *Proc. IEEE 7th Top. Meet. Electr. Performance Electron. Packag.*, 132, 1998.
31. Iqbal, A., Swaminathan, M., Nealon, M., and Omer, A., Design trade-offs among MCM-C, MCM-D and MCM-D/C technologies, *Proc. 1993 IEEE Multi-Chip Module Conf.*, 12, 1993.
32. Thompson, P., MCM-L product development process for low-cost MCMs, *Proc. 1994 Int. Conf. Multichip Modules*, 449, 1994.
33. Begay, M.J. and Cantwell, R., MCM-L cost model and application case study, *Proc. 1994 Int. Conf. Multichip Modules*, 332, 1994.
34. Cokely, D. and Strittmatter, C., Redefining the economics of MCM applications, *Proc. 1994 Int. Conf. Multichip Modules*, 306, 1994.
35. Pecht, M.G., Agarwal, R., McCluskey, P., Dishongh, T., Javadpour, S., and Mahajan, R., *Electronic Packaging Materials and Their Properties*, CRC Press, Boca Raton, FL, 1999.
36. Andrews, J., Mahalingam, L., and Berg, H., Thermal characteristics of 16- and 40-pin plastic DIPs, *IEEE Trans. Components, Hybrids, Manuf. Technol.*, 4, 455, 1981.
37. Mulgaonker, S., Chambers, B., and Mahalingam, M., An assessment of the thermal performance of the PBGA family, *11th IEEE Semicond. Thermal Meas. Manage. Symp.*, 17, 1995.
38. Edwards, D., Hwang, M., and Stearns, B., Thermal enhancement of IC packages, *Proc. 10th IEEE/CPMT Semicond. Thermal Meas. Manage. Symp.*, 33, 1994.
39. Gallo, A.A. and Munamarty, R., Popcorning: a failure mechanism in plastic-encapsulated microcircuits, *IEEE Trans. Reliability*, 44, 362, 1995.
40. Ahn, S.-H. and Kwon, Y.-S., Popcorn phenomena in a ball grid array package, *IEEE Trans. Components, Packag. Manuf. Technol., Part B: Adv. Packag.*, 18, 491, 1995.
41. Gannamani, R. and Pecht, M., An experimental study of popcorning in plastic encapsulated microcircuits, *IEEE Trans. Components, Packag. Manuf. Technol*, Part A, 19, 194, 1996.
42. Alpern, P., Lee, K.C., Dudek, R., and Tilgner, R., A simple model for the mode I popcorn effect for IC packages with copper leadframe, *IEEE Trans. Components Packag. Technol.*, 25, 301, 2002.
43. Alpern, P., Dudek, R., Schmidt, R., Wicher, V., and Tilgner, R., On the mode II popcorn effect in thin packages, *IEEE Trans. Components Packag. Technol.*, 25, 56, 2002.
44. Pecht, M., *Integrated Circuit, Hybrid and Multichip Module Package Design Guidelines: A Focus on Reliability*, John Wiley & Sons, New York, 1994.
45. Ghandhi, S.K., *VLSI Fabrication Principles*, 2nd ed., John Wiley & Sons, New York, 1994.
46. Ghandhi, S.K., *The Theory and Practice of Microelectronics*, Robert E. Krieger, Malabar, FL, 1968.

APPENDIX A

Properties of Si and GaAs at 300 K

Properties	Si	GaAs
Atomic density (cm^{-3})	5.0×10^{22}	4.42×10^{22}
Breakdown field (V/cm)	3×10^{5}	4×10^{5}
Crystal structure	Diamond	Zinc blende
Density (g/cm^{3})	2.328	5.32
Dielectric Constant, ε_S	11.9	13.1
Effective density of states in the conduction band, N_C (cm^{-3})	2.8×10^{19}	4.7×10^{17}
Effective density of states in the valence band, N_V (cm^{-3})	1.04×10^{19}	7.0×10^{18}
Energy gap, E_g	1.12	1.424
Intrinsic carrier concentration, n_i (cm^{-3})	1.45×10	1.79×10^{6}
Lattice constant (Å)	5.43095	5.6534
Mobility ($cm^2V^{-1}s^{-1}$)	1500	8500
	450	400
Thermal conductivity ($Wcm^{-1}/°C$)	1.50	0.46

APPENDIX B

Design Rules, Constants, Symbols, and Definitions

B.1 Design Rules

$\mu_n = 580\ cm^2V^{-1}s^{-1}$	Surface electron mobility for silicon n-MOSFETs
$\mu_p = 230\ cm^2V^{-1}s^{-1}$	Surface hole mobility for silicon p-MOSFETs
$V_D = 0.7\ V$	p–n diode turn-on voltage
$V_{SBD} = 0.3\ V$	Schottky diode turn-on voltage
$V_{BEA} = 0.7\ V$	Base–emitter forward active voltage
$V_{BES} = 0.8\ V$	Base–emitter saturation voltage
$V_{CES} = 0.1\ V$	Collector–emitter saturation voltage
$V_{BEOH} = 0.8\ V$	Base–emitter on-hard voltage
$V_{CEOH} = 0.5\ V$	Collector–emitter on-hard voltage

B.2 Constants

ε_0	$8.85 \times 10^{-14}\ F/cm$	Permittivity of free space
ε_{ox}	$3.9\varepsilon_0$	Permittivity of silicon dioxide
ε_{Si}	$11.9\varepsilon_0$	Permittivity of silicon
k	$1.28066 \times 10^{-23}\ J/K$	Boltzmann constant
q	$1.602 \times 10^{-19}\ C$	Electronic charge

B.3 Symbols

α	Switching activity factor (NMOS, CMOS)
α_F	Forward active common-base current gain (BJT)
α_R	Reverse active common-base current gain (BJT)

α_T	Base transport factor (BJT)
β_F	Forward active common-emitter current gain (BJT)
β_R	Reverse active common-emitter current gain (BJT)
ΔL	Reduction in channel length (MOSFET)
ε	Permittivity
ε_0	Permittivity of free space
ε_{DI}	Permittivity of the dielectric (interconnect)
ε_{ox}	Permittivity of silicon dioxide
ε_{Si}	Permittivity of silicon
ϕ_{MS}	Metal–semiconductor work function difference (MOSFET)
γ	Body effect coefficient (MOSFET)
γ_E	Emitter injection efficiency (BJT)
λ	Optical wavelength
$\lambda(t)$	Instantaneous failure rate
μ_{DI}	Permeability of the dielectric (interconnect)
μ_n	Mobility (electrons)
μ_p	Mobility (holes)
σ	Conductivity
τ_F	Effective forward lifetime (p–n junction, BJT)
τ_n	Lifetime (electrons)
τ_p	Lifetime (holes)
τ_R	Effective reverse lifetime (p–n junction, BJT)
τ_{SC}	Space charge lifetime (p–n junction)
A	Chip area
A	Junction area
BV	Collector–base breakdown voltage (BJT)
c	Capacitance per unit length (interconnect)
C	Capacitance (interconnect)
C_{dm}	Maximum depletion layer capacitance (MOSFET)
C_{GD}	Gate–drain capacitance (MOSFET)
C_{GS}	Gate–source capacitance (MOSFET)
$C_{interconnect}$	Interconnect capacitance
C_L	Load capacitance
C_{ox}	Oxide capacitance
C_T	Transition capacitance, or depletion capacitance (p–n junction)
D_0	Areal density of defects
D_n	Diffusivity (electrons)
D_{nB}	Diffusivity for electrons in the base (BJT)
D_p	Diffusivity (holes)
D_{pE}	Diffusivity for holes in the emitter (BJT)
E	Electric field intensity
E_A	Activation energy
E_c	Edge of the conduction band
E_f	Fermi level
E_g	Energy gap
E_i	Intrinsic Fermi level

E_v	Edge of the valence band
f	Switching frequency
f_M	Ring oscillator frequency with M inverters
G_0	Generation rate for electron–hole pairs
I_B	Base current (BJT)
I_C	Collector current (BJT)
I_{CC}	Collector supply current
I_{CCH}	Collector supply current with the output high
I_{CCL}	Collector supply current with the output low
I_D	Drain current (MOSFET)
I_{DD}	Drain supply current
I_{DDH}	Drain supply current with the output high
I_{DDL}	Drain supply current with the output low
I_E	Emitter current (BJT)
I_{EE}	Emitter supply current
I_{gen}	Generation current (p–n junction)
I_{IH}	Input high current
I_{IL}	Input low current
I_{nC}	Collector electron current (BJT)
I_{nE}	Emitter electron current (BJT)
I_{OH}	Output high current
I_{OL}	Output low current
I_{pE}	Emitter hole current (BJT)
I_{PN}	p–n junction leakage current (CMOS)
I_S	Reverse saturation current (p–n junction)
$I_{subthreshold}$	Subthreshold current (MOSFET)
J	Current density
J_{switch}	Switching energy (CMOS)
k	Number of emitters (multi-emitter bipolar transistor)
k	Boltzmann constant (diode equation)
k'	Process transconductance parameter (MOSFET)
k_N'	Process transconductance parameter (n-MOSFET)
K_P'	Process transconductance parameter (p-MOSFET)
K	Device transconductance parameter (MOSFET)
K_N	Device transconductance parameter (n-MOSFET)
K_P	Device transconductance parameter (p-MOSFET)
l	Inductance per unit length (interconnect)
L	Gate length (MOSFET)
L	Inductance (interconnect)
L_n	Diffusion length (electrons)
L_N	Gate length (n-MOSFET)
L_{nB}	Diffusion length for electrons in the base (BJT)
L_p	Diffusion length (holes)
L_P	Gate length (p-MOSFET)
L_{pE}	Diffusion length for holes in the emitter (BJT)
LS	Logic swing

m	Capacitance ratio parameter (MOSFET)
m	Collector multiplication coefficient (BJT)
M	Collector multiplication factor (BJT)
M	Fan-in
M	Number of inverters in a ring oscillator
MTTF	Mean time to failure
n	Electron concentration
$n_{\_}$	Empirical velocity saturation coefficient (electrons or holes)
$\bar{n}$	Equilibrium electron concentration
n'	Excess electron concentration
n_i	Intrinsic carrier concentration (semiconductor)
N	Fan-out
N	Emission coefficient (p–n junction)
N	Number of chips on the wafer
N_a	Acceptor concentration
N_{aB}	Acceptor concentration in the base (BJT)
N_d	Donor concentration
N_{dE}	Donor concentration in the emitter (BJT)
N_C	Effective density of states at the edge of the conduction band
N_D	Number of defects on the wafer
N_G	Number of good die on the wafer
N_{MAX}	Maximum fan-out
N_V	Effective density of states at the edge of the valence band
$p_{\_}$	Hole concentration
$\bar{p}$	Equilibrium hole concentration
p'	Excess hole concentration
P_{AC}	Dynamic dissipation
P_{DC}	DC dissipation
P_L	Output low power dissipation
P_H	Output high power dissipation
P_{pn}	p–n junction leakage power (CMOS)
P_{sc}	Short-circuit power (CMOS)
P_{switch}	Capacitance switching power (CMOS)
$P_{subthreshold}$	Subthreshold power (CMOS)
PDP	Power delay product
$q\chi$	Semiconductor electron affinity (MOSFET)
$q\phi_m$	Metal work function (MOSFET)
Q_B	Excess minority carrier charge in the base (BJT)
Q_B	Semiconductor depletion charge under strong inversion (MOSFET)
Q_i	Inversion layer charge
Q_{II}	Ion-implanted charge (MOSFET)
Q_{ox}	Oxide charge (MOSFET)
r	Resistance per unit length (interconnect)
R	Resistance (interconnect)
R	Recombination rate (minority carriers)

$R(t)$	Probability that a circuit will survive to time t
R_{DN}	Source-drain "on" resistance (n-MOSFET)
S	Subthreshold swing (MOSFET)
S_{MAX}	Maximum number of series-gated transistors (ECL)
s	Scaling factor (MOSFETs, CMOS)
t_F	Fall time
t_{ox}	Oxide thickness
t_{PHL}	Output high-to-low propagation delay
t_{PLH}	Output low-to-high propagation delay
t_R	Rise time
t_S	Saturation delay
t_t	Transit time (MOSFET)
t_{tB}	Base transit time (p–n junction or BJT)
v	Carrier velocity (electron or hole)
V_{BE}	Base–emitter voltage (BJT)
V_{BEA}	Base–emitter voltage, forward active (BJT)
V_{BEOH}	Base–emitter voltage, on-hard (Schottky-clamped BJT)
V_{BES}	Base–emitter voltage, saturation (BJT)
V_{BC}	Base–collector voltage (BJT)
V_{BCA}	Base–collector voltage, reverse active (BJT)
V_{bi}	Built-in voltage (p–n junction)
V_{BS}	Body-source bias voltage (MOSFET)
V_{CE}	Collector–emitter voltage (BJT)
V_{CEOH}	Collector–emitter voltage, on-hard (Schottky-clamped BJT)
V_{CES}	Collector–emitter voltage, saturation (BJT)
V_{CC}	Collector supply voltage
V_D	Turn-on voltage (p–n diode)
V_{DD}	Drain supply voltage
V_{DS}	Gate–source voltage (MOSFET)
V_{EE}	Emitter supply voltage
V_{GS}	Gate–source voltage (MOSFET)
V_{IH}	Input high voltage
V_{IL}	Input low voltage
V_{IN}	Input voltage
V_{NMH}	High noise margin
V_{NML}	Low noise margin
V_{OH}	Output high voltage
V_{OL}	Output low voltage
V_P	Pinch-off voltage (MESFET)
V_{REF}	Reference voltage (ECL)
v_{sat}	Saturation velocity (electron or hole)
V_{SBD}	Turn-on voltage (Schottky diode)
V_{ss}	Steady-state voltage
V_T	Threshold voltage (MOSFET)
V_{TH}	V_T for high-threshold devices (DT CMOS)
V_{TL}	Threshold voltage for pull-up device (NMOS)

V_{TL}	VT for low-threshold devices (DT CMOS)
V_{TN}	Threshold voltage for n-channel MOSFET (CMOS)
V_{TO}	Zero-bias threshold voltage (MOSFET)
V_{TO}	Threshold voltage for pull-down device (NMOS)
V_{TP}	Threshold voltage for p-channel MOSFET (CMOS)
W	Depletion width (p-n junction)
W	Gate width (MOSFET)
W_B	Base width (BJT)
W_B	Base width (p–n junction)
W_D	Drain depletion width (MOSFET)
W_{dm}	Depletion width in silicon under inversion (MOSFET)
W_N	Gate width (n-MOSFET)
W_P	Gate width (p-MOSFET)
W_S	Source depletion width (MOSFET)
x_n	Depletion width in the n-type side (p–n junction)
x_p	Depletion width on the p-type side (p–n junction)
Y	Yield

B.4 Definitions

ALSTTL	Advanced low-power Schottky TTL
ASIC	Application-specific integrated circuit
ASTTL	Advanced Schottky TTL
BCC	Body-centered cubic
BGA	Ball grid array
BiCMOS	Bipolar CMOS logic
BJT	Bipolar junction transistor
BOX	Buried oxide
BSIM	Berkeley short-channel IGFET model
CMOS	Complementary metal oxide–semiconductor logic
CMP	Chem-mechanical polishing
CVD	Chemical vapor deposition
DCFL	Direct-coupled FET logic
DDE	Double-diffused epitaxial process (bipolar transistors)
DFT	Design for test
DIP	Dual in-line package
D-MESFET	Depletion MESFET
DT CMOS	Dual threshold CMOS
DTI	Deep trench isolation
DTL	Diode–transistor logic
DVS	Dynamic voltage scaling
DW	Device wafer (wafer-bonded SOI)
ECL	Emitter-coupled logic

ELTRAN	Epitaxial layer transfer (wafer-bonded SOI)
E-MESFET	Enhancement MESFET
FCC	Face-centered cubic
FD	Fully depleted (SOI MOSFET)
FET	Field effect transistor
HSPICE	Synopsys™ version of SPICE
HW	Handle wafer
IGFET	Insulated gate field-effect transistor
I^2L	Integrated injection logic
I/O	Input/output
ITOX	Internal thermal oxidation
ITRS	International technology roadmap for semiconductors
JFET	Junction field-effect transistor
LOCOS	Local oxidation of silicon
LSI	Large-scale integration
LSTTL	Low-power Schottky TTL
MESFET	Metal–semiconductor field-effect transistor
MOSFET	Metal oxide–semiconductor field-effect transistor
MSI	Medium-scale integration
MT CMOS	Multiple-threshold CMOS
MTTF	Mean time to failure
NMOS	n-channel metal oxide–semiconductor logic
PD	Partially depleted (SOI MOSFET)
PDIP	Plastic dual in-line package
PDP	Power delay product
PECL	Positive emitter-coupled logic
PGA	Pin grid array
PSPICE	Cadence™ version of SPICE
QFP	Quad flat pack
RTL	Resistor–transistor logic
SC	Simple cubic
SIMOX	Separation by implantation of oxygen
SOI	Silicon on insulator
SOS	Silicon on sapphire
SPICE	Simulation program with integrated circuit emphasis
SSI	Small-scale integration
STI	Shallow trench isolation
STO	Shallow trench oxide
STTL	Schottky TTL
TTL	Transistor–transistor logic
VHSIC	Very high-speed integrated circuit
VLSI	Very large-scale integration
VPE	Vapor phase epitaxy
VT CMOS	Variable threshold CMOS
WB	Wafer bonded (SOI)

Index

A

B

C

D

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

U

V

W

X

Y

Z