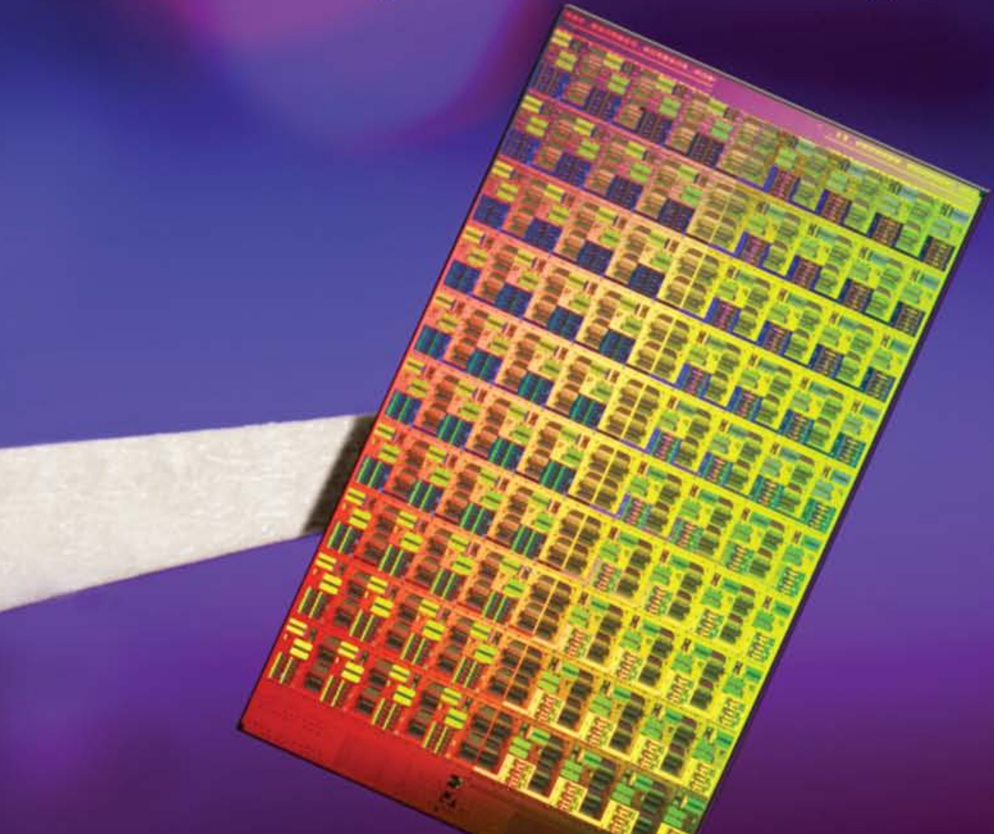
SECOND EDITION

Digital Integrated CIRCUITS

Analysis and Design



John E. Ayers

CRC Press
Taylor & Francis Group

SECOND EDITION

Digital Integrated CIRCUITS

Analysis and Design

SECOND EDITION

Digital Integrated CIRCUITS

Analysis and Design

John E. Ayers

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

No claim to original U.S. Government works
Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number-13: 978-1-4200-6987-7 (Hardcover)

This book contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Library of Congress Cataloging-in-Publication Data

Ayers, John E.
Digital integrated circuits : analysis and design / by John E. Ayers. -- 2nd ed.
p. cm.
Includes bibliographical references and index.
ISBN-13: 978-1-4200-6987-7
ISBN-10: 1-4200-6987-X
1. Digital integrated circuits--Design and construction. I. Title.

TK7874.65.A94 2010
621.3815--dc22 2009010227

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

To Kimberly, Jacob, Sarah, and Rachel.

Their patience and limitless support

made this project possible.

Contents

Preface

Today, there is no field of enterprise more dynamic or challenging than 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 18–24 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 unmatched in any endeavor of mankind and has revolutionized the way we live and work.

The rapid progress in digital circuitry, and the relentless scaling of MOS transistors that brought it about, have broadened the interdisciplinary reach of the field. More than ever, the materials, processing, device physics, and circuit performance characteristics are inseparably linked. The original book, conceived in this interdisciplinary spirit and with a strong focus on principles, bridged a void that had existed between books on transistor electronics and those treating VLSI design and fabrication as a separate topic.

This second edition adheres to this same successful concept but is significantly improved in nearly every way, with four new chapters, more than 200 new illustrations, and support provided on a dynamic website (http://www.engr.uconn.edu/ece/books/ayers), including a section for instructors. An all-new chapter describes the integrated circuit fabrication process with detailed illustrations and discussions of the dual damascene process for copper interconnect, metal gates, and high-κ gate dielectric. The expanded chapter on the MOS transistor includes new material on the physics of short-channel devices. CMOS circuitry has been covered in greater detail by incorporating numerous new examples and short-channel behavior.

Like the first edition, this book bridges the gap between courses in transistor electronics and VLSI design or fabrication. It serves as a crucial link for integrated circuit engineers, because they make the cross-disciplinary connections to guide them in more advanced work. For pedagogical reasons, this book uses SPICE level 1 models (similar to the transistor electronics courses) but introduces BSIM models that are indispensable for VLSI design. This approach makes it possible to draw direct connections between the hand analysis and the SPICE models for the development of a strong and intuitive sense of device and circuit design. Once these connections are made, the BSIM device models can be better appreciated as incorporating many second-order and empirical corrections to the predictions of the level 1 model.

Digital Integrated Circuits, Second Edition, focusing on principles and presented from a modern interdisciplinary view, should serve integrated circuits engineers from all disciplines for years to come.

John E. Ayers
Storrs, Connecticut

About the Author

John E. Ayers grew up eight miles from an integrated circuit design and fabrication facility, where he worked as a technician and first developed his passionate interest in the topic. After earning a BSEE degree from the University of Maine (Orono, Maine) in 1984, he worked as an integrated circuit test engineer for National Semiconductor (South Portland, Maine). He worked for six years at Rensselaer Polytechnic Institute (Troy, New York) and Philips Laboratories (Briarcliff, New York) on semiconductor material growth and characterization, earning the MSEE in 1987 and the PhDEE in 1990, both from Rensselaer Polytechnic Institute. Since then, he has been employed in academic research and teaching at the University of Connecticut (Storrs, Connecticut), where he has taught the course on digital integrated circuits for a number of years. He has been honored with the Electrical and Computer Engineering Best Teacher Award (2003–2004 and 2004–2005) and the School of Engineering Outstanding Teaching Award (2000–2001) and is a University of Connecticut Teaching Fellow (1999–2000). Ayers has authored more than 60 journal and conference papers as well as three books. He is a member of Eta Kappa Nu, Tau Beta Pi, and Phi Kappa Phi and is a senior member of the Institute of Electrical and Electronics Engineers. He lives in Ashford, Connecticut, and enjoys running, hiking, and bicycling with his wife and three children.

1

Introduction

1.1 Historical Perspective and Moore's Law

Since the demonstration of the first transistor [1–4] at Bell Laboratories in 1947 (Figure 1.1), rapid progress in the design and manufacture of semiconductors has led to gigahertz microprocessors and gigabit memories today. The invention of the integrated circuit (Figure 1.2) in 1958 [5, 6], and subsequent improvements on the concept in the 1960s, made it possible to combine multiple devices on a single chip of silicon instead of wiring together devices on a circuit board. Thus, it was possible to reduce the size, weight, and cost of a circuit even while increasing its functionality. These breakthroughs eventually allowed the fabrication of billions of transistors in a single chip of silicon, enabling computing power far

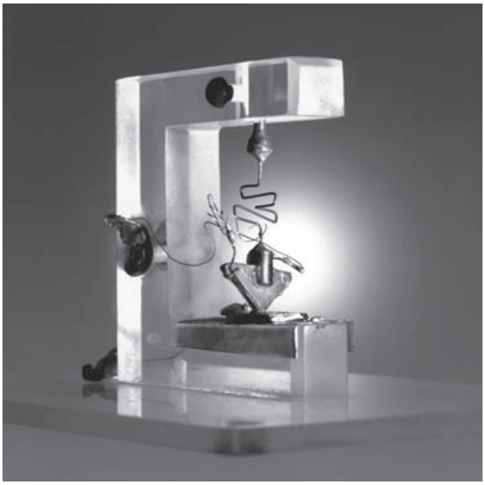


FIGURE 1.1
The first transistor, invented at Bell Laboratories in 1947. (From Lucent Technologies Inc./Bell Labs. With permission.)

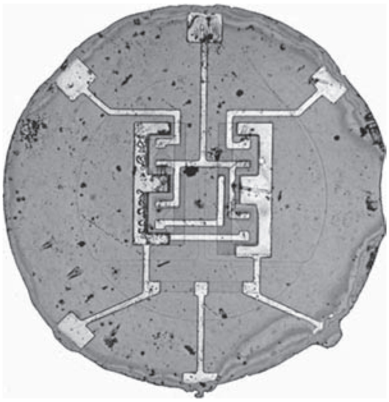


FIGURE 1.2
The first commercial integrated circuit, developed by Robert Noyce at Fairchild Semiconductor. The integrated circuit was invented by Robert Noyce of Fairchild Semiconductor and Jack Kilby of Texas Instruments at about the same time. (From Fairchild Semiconductor. With permission.)

beyond that achievable by wiring together discrete transistors. At the present time, state-of-the-art microprocessor chips contain hundreds of millions of transistors, whereas memory chips contain up to several billion transistors.

Soon after the realization of integrated circuits, Intel cofounder Gordon Moore noted that the complexity of integrated circuits was increasing exponentially with time. Moore stated that the "complexity" of integrated circuits with "minimum component costs" was doubling every 12 months [7]. Since that time, several variations of "Moore's law" have been stated. Usually, it is noted that the number of transistors per microprocessor chip doubles every 24 months, whereas the number of bits per dynamic random access memory (DRAM) chip doubles every 18 months. Remarkably, this rate of progress has been maintained for more than three decades! Figure 1.3 illustrates this exponential progress for microprocessors and DRAMs. Similar exponential trends have been established for flash memories and application-specific integrated circuits (ASICs). However, memory chips have outpaced microprocessors because of their simpler designs and built-in redundancy.

Industry has kept pace with Moore's law primarily by scaling down the dimensions of transistors through improved lithography [8, 9], but chip size has also been increased. Transistor miniaturization has been pursued aggressively and has brought about improvements in circuit performance and cost as well as density. A key parameter describing this scaling is the "minimum feature size" for the transfer of circuit patterns from a computer design to the semiconductor wafer. Figure 1.4 shows the historical evolution

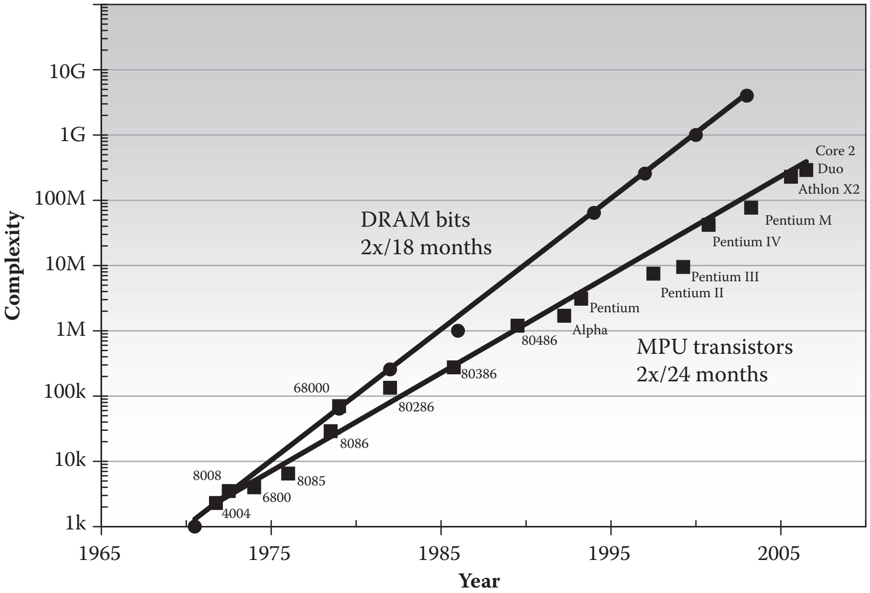


FIGURE 1.3
According to Moore's law, the complexity of integrated circuits has increased exponentially with time. Historically, the number of transistors per microprocessor unit (MPU) doubles every 24 months; the number of bits per DRAM chip doubles every 18 months.

of the minimum feature size, along with projections out to the year 2020. Historically, much of the progress shown in Figure 1.4 was achieved by decreasing the wavelength used in lithographic tools. However, at the present time, technical solutions for the lithographic requirements of 2020 and beyond have not been found.

With the goal of extending the historic trends in integrated circuit technology, the Semiconductor Industry Association in the United States [10] produced the National Technology Roadmap for Semiconductors in 1992. This roadmap defined industry-wide technology goals with a 15-year horizon and was revised in 1994 and 1997. In 1998, after 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 [11]. Full ITRS reports are published biannually, in odd years, and update reports are published in even years.

Each ITRS report projects industry trends 15 years into the future. What, then, is the digital integrated circuit industry expected to look like in 2020? According to the 2007 ITRS, the starting silicon wafers will grow to 450 mm in diameter, whereas transistor gate lengths will diminish to 5.6 nm. As a consequence of these developments, flash memory will be able to

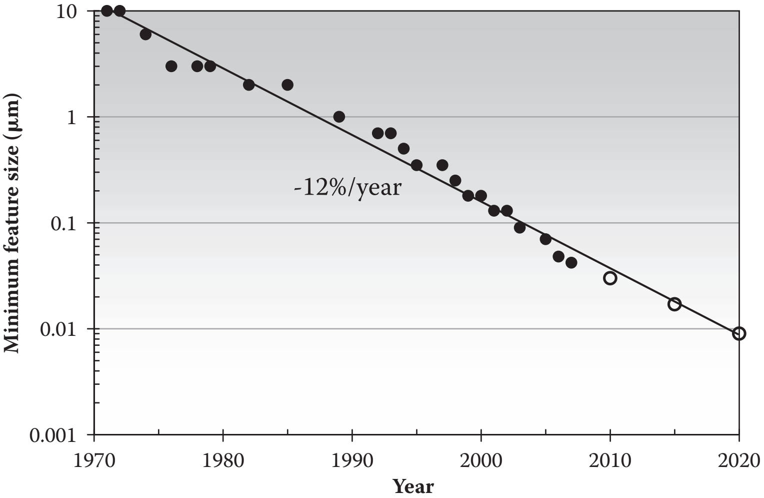


FIGURE 1.4
Minimum feature size for the fabrication of integrated circuits as a function of time.

store one terabit on a single chip, and you will be able to purchase a 12.4 GHz processor with 6.1 billion transistors and 7902 pins for about 0.24 microcents per transistor! These and other important trends are charted in Table 1.1.

The rapid progress in digital integrated circuits has been fueled by strong demand from a number of driving force applications. During the first decade after the invention of the integrated circuit, mainframe computers represented the driving application. In the 1980s, there was a shift to personal computers, consumer electronics, and digital communication. During the 1990s, wireless communication, portable computers, and handheld devices came into play. Today, video-on-demand, multimedia applications, and network computing are some of the driving forces behind the technology. Sales of semiconductors have increased at a 14.9% compound annual growth rate since the invention of the integrated circuit (see Figure 1.5), now making up 20% of the world electronics market and 2% of the gross world product (GWP).

At the present time, most digital integrated circuits use the CMOS circuit family, so named because it uses complementary metal oxide semiconductor transistors (MOS transistors, or MOSFETs). However, this hasn't always been the case. The first integrated circuit was based on bipolar junction transistors (BJTs), and numerous families of digital integrated circuits have been realized using BJTs; these include resistor transistor logic, diode transistor logic, transistor transistor logic (TTL), integrated injection logic, and emitter coupled logic (ECL). New bipolar circuit families were still being developed in the 1980s, because BJTs provided a speed advantage over metal oxide semiconductor field effect transistors (MOSFETs) using the lithographic design

TABLE 1.1

Semiconductor Technology Trends

Year of production	2007	2010	2015	2020
Lithography				
DRAM stagger-contacted metal ½ pitch (nm)*	65	45	25	14
MPU/ASIC stagger-contacted metal ½ pitch (nm)*	68	45	25	14
Flash uncontacted poly Si ½ pitch*	54	36	20	11
MPU printed gate length (nm)	42	30	17	9
MPU physical gate length (nm)	25	18	10	5.6
# mask levels, MPU	33	35	37	39
# mask levels, DRAM	24	26	26	26
Maximum number of wiring levels	11	12	13	14
MPU high performance				
MPU transistors per chip (millions)	386	773	1546	6184
MPU chip size (mm^2)	140	140	88	111
MPU cost per transistor (microcents) (high performance)	12.2	4.3	0.76	0.24
MPU total package pins (high performance)	4000	4851	6191	7902
Clock frequency (GHz)	4.70	5.875	8.522	12.361
DRAM				
DRAM bits per chip (billions)	2G	4G	8G	32G
DRAM chip size (mm^2)	93	93	59	74
DRAM cost per bit (microcents)	0.96	0.34	0.06	0.01
Flash memory				
Flash bits per chip (four level cell)	32G	64G	128G	1024G
Flash chip size (mm^2)	144	128	81	102
ASICs				
ASIC package pins (high performance)	4000	4851	6191	7902
General				
On-chip clock frequency (GHz)	4.70	5.875	8.522	12.361
Supply voltage V_{DD} (V) (high performance)	1.1	1.0	0.80	0.65
Chip power dissipation (W) (high performance)	189	198	198	198
Silicon wafer diameter (mm)	300	300	300/450	450

Source: The International Technology Roadmap for Semiconductors, http://public.itrs.net.

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

rules in place at that time. By the 1990s, CMOS had become the dominant technology, but bipolar circuits, and also bipolar-CMOS (BiCMOS) circuits, remained commercially important into the beginning of the 21st century.

The MOSFET was invented in 1930 by Lilienfeld [12], but the first working device was not demonstrated until 1960 by Kahng and Atalla (see Figure 1.6) [13, 14]. MOS transistors allowed the integration of more functionality on an integrated circuit than BJTs for two reasons. First, MOSFETs are

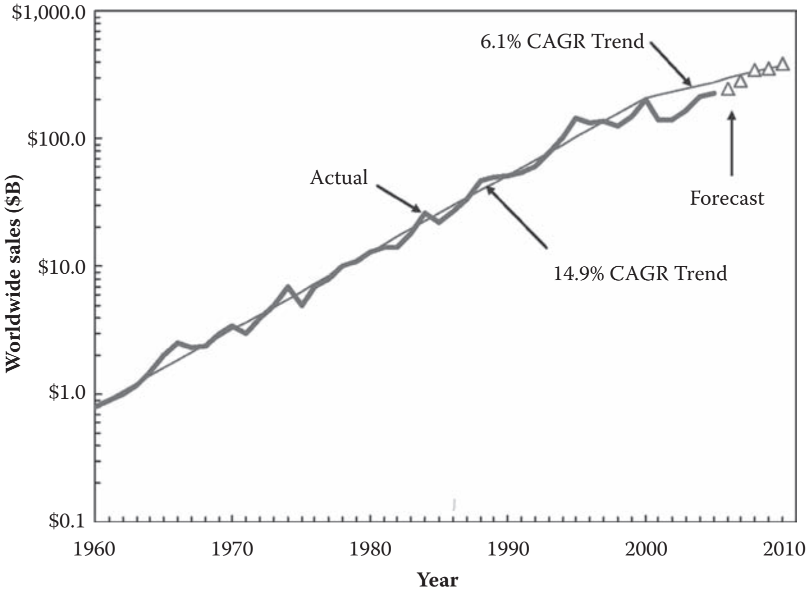


FIGURE 1.5
Worldwide sales of integrated circuits have increased with a 14.9% compound annual growth rate since 1960 and now account for 2% of the gross world product. (From Kumar, R., *IEEE Solid-State Circuits* 12, 22–27, 2007. With permission.)

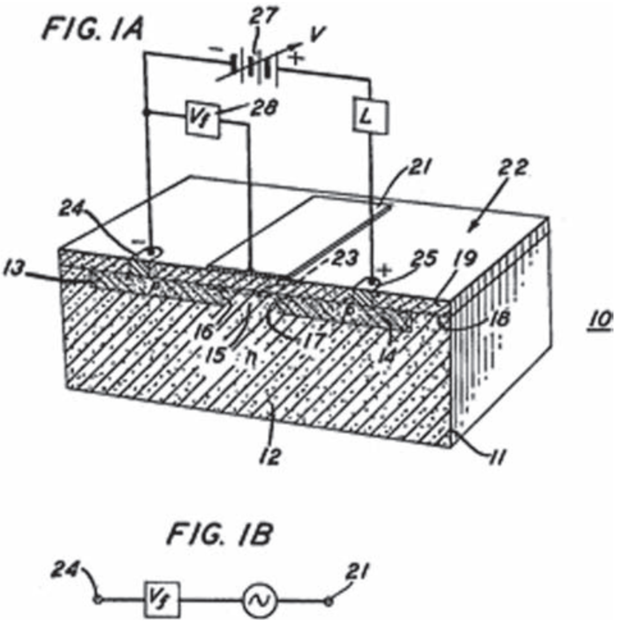


FIGURE 1.6
The first MOSFET. Electric field controlled semiconductor device. (From Kahng, D., and Atalla, M.M., U.S. Patent 3,102,230, filed May 31, 1960, and issued Aug. 27, 1963.)

inherently smaller than BJTs using the same set of layout design rules. Second, MOSFETs are voltage controlled and require no biasing resistors like BJTs, which are current controlled. Integrated resistors take up considerable chip area, because they are typically much larger than either BJTs or MOSFETs.

The first microprocessor was the 4004, introduced by Intel in 1971 (see Figure 1.7). It was realized using the PMOS circuit family and included 2300 p-channel MOSFETs. PMOS was used because, at that time, the repeatable manufacture of enhancement type (normally off) n-channel MOSFETs was not possible because of oxide and contamination-related problems. For these reasons, PMOS was used for microprocessors and their associated components such as memories and peripheral interfaces for a short time.

By 1974, with the problems of n-channel transistors solved, PMOS circuits were replaced by superior NMOS (n-channel MOSFETs) circuits. Notable examples were the 6800 and 8080 microprocessors.

Soon thereafter, the ability to manufacture both n-MOS and p-MOS transistors on the same wafer enabled the development of CMOS digital integrated circuits. The key advantage of this circuit family, relative to PMOS and NMOS, is its low static power dissipation. After the invention of CMOS in 1963, the 4000 series of CMOS logic gates was developed in 1968, and CMOS microprocessors emerged in the mid-1970s.

Even after the establishment of CMOS as a mainstream digital integrated circuit technology, bipolar circuits temporarily maintained their edge in critical high-speed applications. This is because bipolar transistors exhibited much higher transconductance values than MOSFETs, for a given device size, and could therefore drive high-capacitance off-chip loads at higher data rates. Bipolar circuit families, such as ECL and Schottky transistor-transistor logic (STTL), found use in high-performance computing, supercomputers, and high bit rate communication. Even nonsilicon integrated circuits, such as direct coupled FET logic circuits

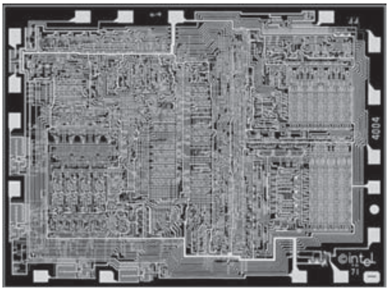


FIGURE 1.7
The Intel 4004 microprocessor. (From Intel. With permission.)

fabricated in gallium arsenide or indium phosphide, were used in high-end microprocessor and digital communication applications. Eventually, however, the relentless scaling of transistor dimensions helped CMOS overtake bipolar circuit families in terms of on-chip switching speed *and* off-chip data rates.

There are limits to the scaling and performance of CMOS circuits, however. At some point in time, it will be necessary to make the transition to a new, higher-density, higher-performance circuit family. We cannot be sure when this will happen or what types of devices might replace MOSFETs, but it is sure to happen. It is important to study digital integrated circuits in this light, with an emphasis on general principles, but the examples are oriented to the CMOS circuit family.

1.2 Electrical Properties of Digital Integrated Circuits

In digital circuitry, voltage signals take on one of two (or possibly more) discrete levels. This contrasts with the case of analog circuits and systems, in which 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 use semiconductor electronic devices to process or combine binary signals in a desired manner. 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 (or some other semiconductor crystal). The level of integration is classified according to the number of gates that have been integrated on a single chip. The various levels of integration have been called small scale integration (SSI), medium scale integration (MSI), large scale integration (LSI), very large scale integration (VLSI), and ultra large scale integration (ULSI) and are listed in Table 1.2. At the present time, integrated circuits at all of these levels of complexity are manufactured and used for various applications. Often, the distinction between VLSI and ULSI is not made, so that state-of-the art circuits are often referred to as "VLSI" even today.

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 because the size of silicon wafers grew to 200 mm and then 300 mm. Nonetheless, it has become feasible to implement "system on a chip" designs, in which an entire

TABLE 1.2
Levels of Integration

Level of integration	Gates/chip
Small scale integration	1–10
Medium scale integration	10–100
Large scale integration	100–10^4
Very large scale integration	>10^4

computer system is built in a single chip of silicon. It goes without saying that this approach is superior in size, cost, and performance compared with the traditional approach of wiring together many integrated circuits on a printed circuit board.

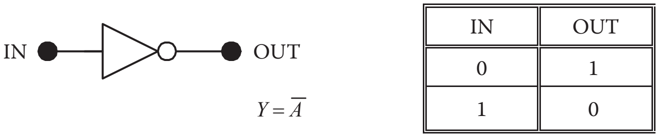
In this section, we will describe the electrical properties of digital integrated circuits at the gate level. Ideally, a logic gate should process an infinite number of inputs, perform some logic function with zero time delay, be completely immune to the effects of loading by other gates, and consume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

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. In this number system, each binary digit, or bit, takes on a value of "0" or "1." Sometimes a 1 is referred to as a "true" result, whereas a 0 is referred to as a "false" result.

A commonly used logic gate is the one-input inverter, or NOT gate, shown in Figure 1.8. If the input A is true, then the output Y is not true and vice versa. The implementation of an inverter requires one or more switching devices. Three possible implementations are shown in Figure 1.9. Here, the switches are assumed to be three-terminal devices, but nonlinear two-terminal devices can also act as switches.

The active pull-down implementation of Figure 1.9a uses a single switch device connected between the output and ground. With a low (logic zero)



IN	OUT
0	1
1	0

FIGURE 1.8
Inverter.

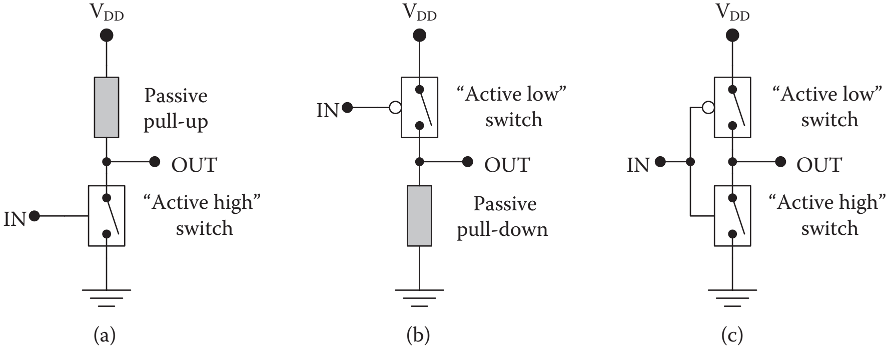


FIGURE 1.9
Inverter circuit designs: (a), Active pull-down design, (b) active pull-up design, and (c) fully active design.

input, the switch is "off" (open) and the output is brought high by the passive pull-up device. (The passive pull-up device can be a resistor, but other devices may be used as well.) With a high (logic one) input, the switch is "on" (closed) and the output goes low. Therefore, the circuit functions as an inverter, and this type of switch is called an "active high" device. The active pull-down circuit design dissipates steady power under the low output condition, in which both the switch and the passive design are conducting.

In the active pull-up design of Figure 1.9b, the switch is connected between the supply voltage V_{DD} and the output, and a passive pull-down device is used. The circuit still functions as an inverter, as long as an "active low" switch is used. Steady power dissipation occurs in this circuit with the high output condition, as a result of the simultaneous conduction of the switch and the passive design.

The fully active inverter design of Figure 1.9c uses two switch devices; the pull-down switch is active high, whereas the pull-up switch has active low character. It is possible to design this circuit so that the two switches do not conduct simultaneously under static conditions, so the average dissipation can be reduced dramatically.

The switch devices in Figure 1.9 may be any three-terminal active devices. If current-controlled devices are used, such as bipolar junction transistors, it is necessary to include biasing circuitry. Simpler implementations stem from the use of voltage-controlled devices such as MOS transistors.

Figure 1.10 shows inverter circuit implementations based on the use of MOSFETs. The active pull-down design of Figure 1.10a uses all n-channel MOSFETs, the active pull-up design of Figure 1.10b uses all p-channel MOSFETs, and the fully active design of Figure 1.10c uses complementary n-channel and p-channel MOSFETs. These three circuit families are referred to as NMOS, PMOS, and CMOS, respectively.

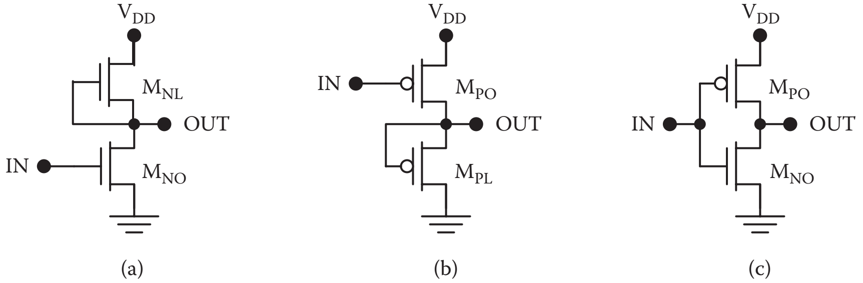
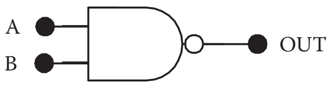


FIGURE 1.10
Inverter circuit implementations using MOSFETs: (a) Active pull-down circuit constructed with n-channel MOSFETs, (b) active pull-up circuit using p-channel MOSFETs, and (c) fully active circuit using complementary MOSFETs.

Other important logic functions include NAND and NOR. These can also be implemented in circuits using active pull down, active pull up, or both. Figure 1.11 shows the NAND (Not AND) gate with its truth table. Three possible circuit designs are provided in Figure 1.12. In the active pull-down circuit of Figure 1.12a, the active high switches are placed in series. Therefore, the output goes low if both inputs are high, causing both switches to turn "on." The active pull-up circuit of Figure 1.12b uses active low switches in parallel. In this case, the output goes high if either input goes low. The fully active circuit of Figure 1.12(c) uses both active high switches in series and active low switches in parallel. Figure 1.13 shows MOSFET-based two-input NAND gates designed using these three basic approaches.

The two-input NOR (Not OR) gate is shown in Figure 1.14 with its truth table. The active pull-down design of Figure 1.15a uses two active high switches in parallel. If either input goes high, the associated switch turns on and brings the output low. The active pull-down design of Figure 1.15b uses two active low switches in series so that, if both inputs go low, the output goes high. The fully active design uses active high pull-down devices and active low pull-up devices. These three designs may be implemented using



$Y = \overline{AB}$

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

FIGURE 1.11
Two-input NAND gate.

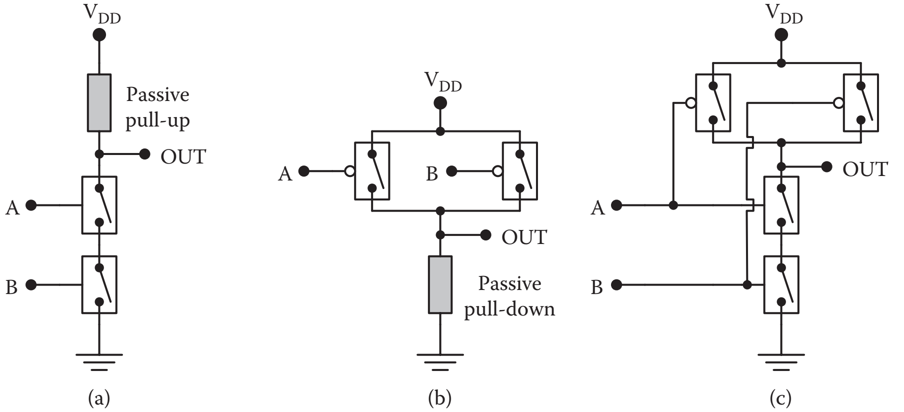


FIGURE 1.12
Design of two-input NAND gates: (a) Active pull-down circuit, (b) active pull-up circuit, and (c) completely active circuit design.

n-channel MOSFETs as the active high devices and p-channel MOSFETs as the active low devices as shown in Figure 1.16.

Another noteworthy circuit is the transmission gate, although it does not perform a Boolean algebraic function; this circuit is shown in Figure 1.17. With logic one applied to the control input, the transmission gate is enabled and the output follows the input. 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

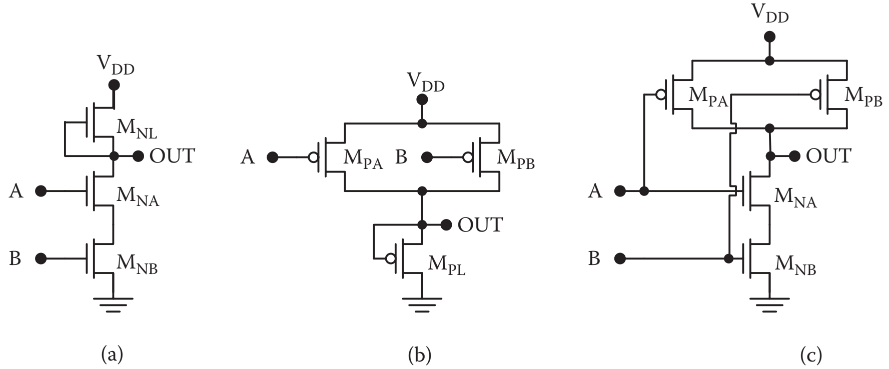
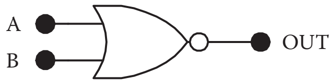


FIGURE 1.13
Two-input NAND circuit implementations using MOSFETs: (a) Active pull-down circuit constructed with n-channel MOSFETs, (b) active pull-up circuit using p-channel MOSFETs, and (c) fully active circuit using complementary MOSFETs.



$Y = \overline{A + B}$

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

FIGURE 1.14
Two-input NOR gate.

circuitry connected to the node. Therefore, transmission gates can be used to connect/disconnect logic blocks in a system. This is useful in bus-based systems and power-managed digital systems. Figure 1.18 illustrates possible designs for the transmission gate, and MOS circuit implementations are given in Figure 1.19.

Practical digital systems involve complex functions of many inputs. However, these complex functions can be realized using the basic functions described above. In fact, it is possible to realize any arbitrary logic function with any arbitrary number of inputs using just the NOT and OR functions or just the NOT and AND functions. There are a number of techniques available for the simplification of complex logic functions that make it possible to realize the necessary logic functions with maximum efficiency.

It is important to note that the circuits described in this section are all *static* logic gates. That is, they are designed to operate with steady (static) input and output voltages. *Dynamic* logic gates are also of tremendous importance in

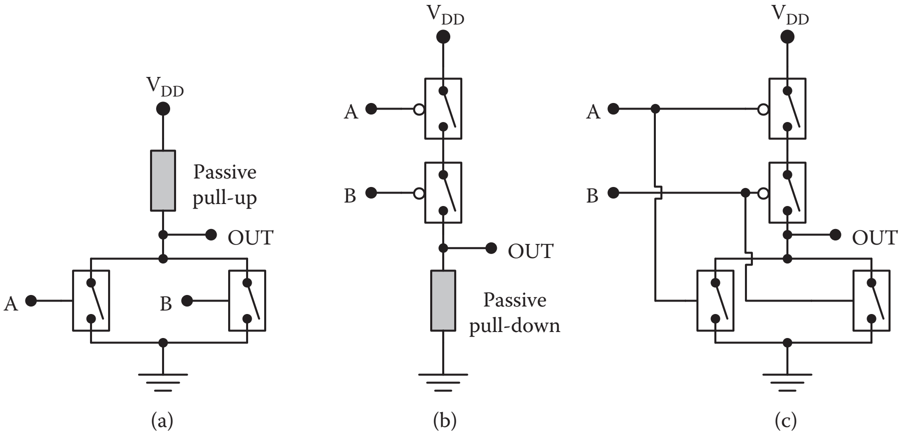


FIGURE 1.15
Design of two-input NOR gates: (a) Active pull-down design, (b) active pull-up design, and (c) fully active design.

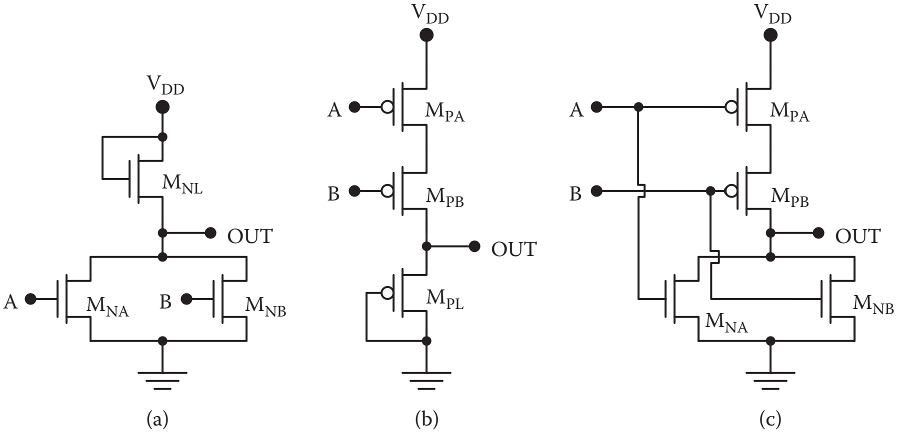


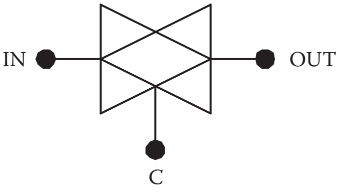
FIGURE 1.16
Two-input NOR circuit implementations using MOSFETs: (a) Active pull-down circuit constructed with n-channel MOSFETs, (b) active pull-up circuit using p-channel MOSFETs, and (c) fully active circuit using complementary MOSFETs.

modern VLSI circuits. This type of gate circuit is controlled by a clock signal, and the output must be evaluated at particular points in the clock cycle. The principles underlying dynamic logic gates will be described in Chapter 8.

1.2.2 Static Voltage Transfer Characteristics

An important electrical characteristic of any static logic gate is the *voltage transfer characteristic* (VTC). This is the steady-state output voltage versus input voltage characteristic. It is usually measured under low-frequency, quasi-static conditions.

The important features of the VTC can be seen in Figure 1.20, which is a generic characteristic for an inverter. There are five critical voltages for the inverter: V_{OL}, V_{OH}, V_{IL}, V_{IH}, and V_M.



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

FIGURE 1.17
Transmission gate.

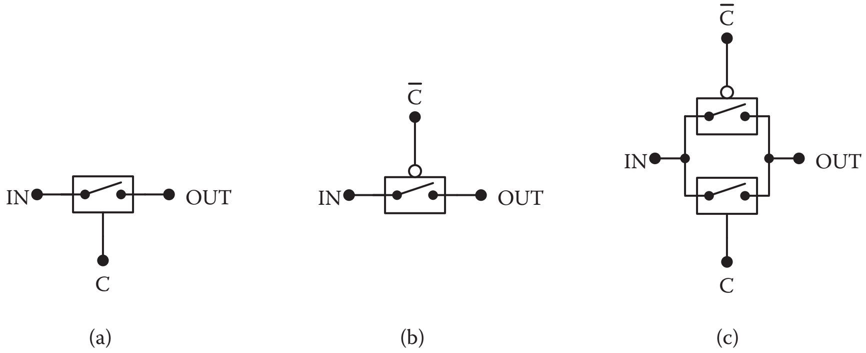


FIGURE 1.18
Design of transmission gates using three-terminal switch devices: (a) Active high design, (b) active low design, and (c) active high-active low design.

The *output low voltage* V_{OL} is the voltage output corresponding to logic zero, and the *output high voltage* V_{OH} is the value of the output corresponding to logic one. The difference between the two output levels is called the *logic swing* LS:

$$\text{LS} = V_{OH} - V_{OL}. \quad (1.1)$$

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 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 voltage

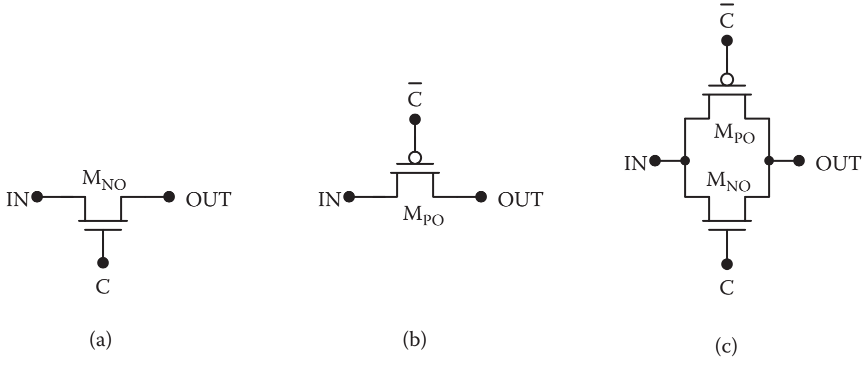


FIGURE 1.19
MOS transmission gate circuits: (a) Active high design using an n-channel MOSFET, (b) active low design using a p-channel MOSFET, and (c) active high-active low design using complementary MOSFETs.

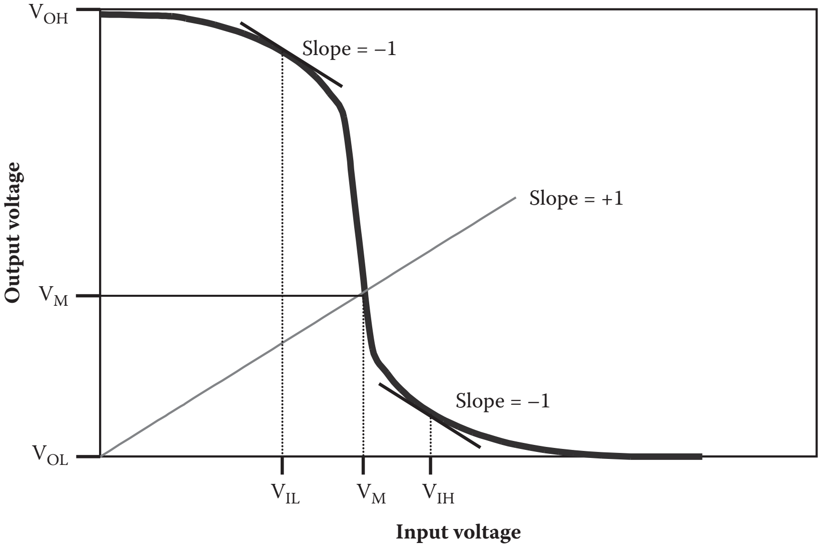


FIGURE 1.20
Voltage transfer characteristic for an inverter.

range. By definition, V_{IL} and V_{IH} are the input voltages for which the slope of the transfer characteristic is minus one (plus one for a noninverting gate):

$$\left.\frac{dV_{OUT}}{dV_{IN}}\right|_{V_{IN}=V_{IL}} = -1 \,. \tag{1.2}$$

and

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

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

$$V_{NML} = V_{IL} - V_{OL} \tag{1.4}$$

and

$$V_{NMH} = V_{OH} - V_{IH} \,, \tag{1.5}$$

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.

The midpoint voltage V_M, sometimes also called the switching threshold, is the value of the input voltage for which $V_{OUT} = V_{IN} = V_M$. Ideally, the midpoint voltage should be halfway between the logic zero and logic one input voltages. Therefore, its value is often used in the design of level translators to interface between circuits operating with different supply voltages.

Static 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 an inverter. The situation is more complicated with multiple inputs, and the usual approach is to tie all but one input to logic zero or logic one, thus avoiding the need for a multidimensional plot. For example, with a NAND gate, all inputs are tied to the positive supply voltage except one, and the transfer characteristic is measured for this one input. For an NOR gate, all inputs but one are grounded for the measurement. Usually, it is assumed that all inputs behave in identical manner, although there may be subtle differences as shown in Chapter 6.

For an ideal logic gate, the output high voltage is equal to the positive supply voltage ($V_{OH} = V_{DD}$), and the output low voltage is equal to the negative supply voltage ($V_{OL} = V_{SS}$ and usually $V_{SS} = 0$). 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 ($V_M = \left(V_{DD} + V_{SS}\right) / 2$). Together, these conditions maximize the noise margins.

1.2.3 Transient Characteristics

Transient characteristics are of great importance because of their bearing on the speed characteristics of digital circuits, such as the clock frequency and off-chip data rates.

Consider Figure 1.21, which shows the transient behavior for an inverter with a rectangular wave applied at the input. There are four important transient parameters for the inverter. These 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 propagation delays are measured between the 50% points on the input and output waveforms. By this, we mean 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. The estimation of the propagation delay t_{PLH} may be understood with the aid of Figure 1.22. Here, it has been assumed that a signal

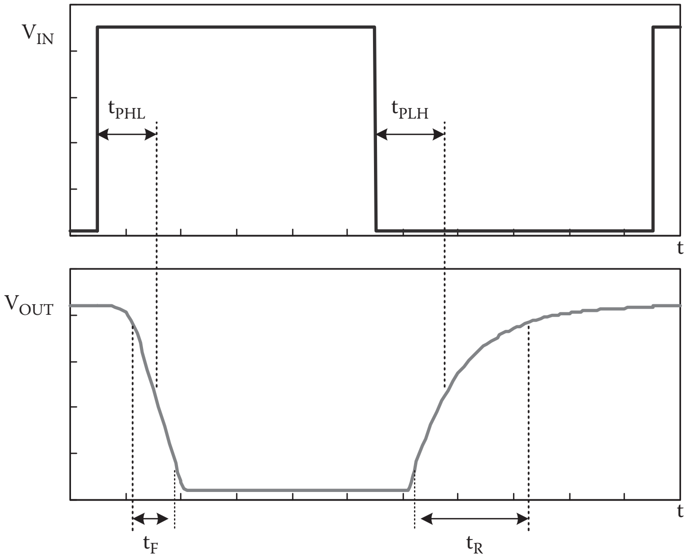


FIGURE 1.21
Transient response of an inverter with a rectangular input waveform.

with an abrupt high-to-low transition has been applied to the input of an inverter with a capacitive load C_L. If the currents in the pull-up and pull-down devices are i_{DD} and i_{SS}, respectively, then t_{PLH} may be found from

$$\frac{1}{2}(V_{OH}+V_{OL})-V_{OL}=\int_0^{t_{PLH}}\frac{i_{DD}-i_{SS}}{C_L}dt\,. \tag{1.6}$$

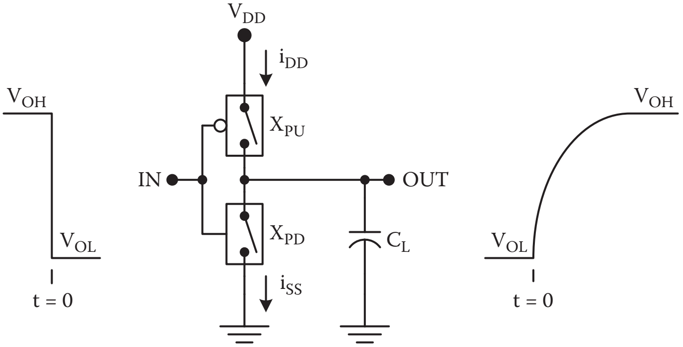


FIGURE 1.22
Determination of the propagation delay t_{PLH} for an inverter.

In general, both i_{DD} and i_{SS} are functions of the output voltage, so they will vary with time. If the pull-down device turns off abruptly at $t = 0$ and the inverter exhibits rail-to-rail swing at the output, then Equation 1.6 simplifies to

$$\frac{V_{DD}}{2} = \frac{1}{C_L} \int_0^{t_{PLH}} i_{DD}\, dt\,. \tag{1.7}$$

The estimation of t_{PHL} is similar. Here it is assumed that the input signal makes an abrupt transition from V_{OL} to V_{OH} at $t = 0$. If the pull-up device switches off abruptly and if the circuit exhibits rail-to-rail swing, then the high-to-low propagation delay can be estimated from

$$\frac{V_{DD}}{2} = \frac{1}{C_L} \int_0^{t_{PHL}} i_{SS}\, dt\,. \tag{1.8}$$

For some digital gate circuits, the two propagation delays t_{PLH} and t_{PHL} are very different so that both should be specified. Often, an average propagation delay t_P is used:

$$t_P = \frac{t_{PLH} + t_{PHL}}{2}\,. \tag{1.9}$$

For a gate circuit with perfectly symmetric transient characteristics, $t_{PLH} = t_{PHL} = t_P$. In practice, the input waveforms to digital gates do not make abrupt transitions but have finite rise and fall times. These are important because they modify the time dependences of the currents i_{DD} and i_{SS} flowing in the gate circuit and therefore the propagation delays. The rise time is measured between the 10 and 90% points on the waveform, and the fall time is measured between the 90 and 10% points. The rise and fall times may be estimated using a similar approach to that applied for the propagation delays.

There is an inverse relationship between the worst-case propagation delay and the maximum achievable system clock frequency:

$$f_{CLK} < \frac{1}{N_S t_P}\,, \tag{1.10}$$

where f_{CLK} is the clock frequency, and N_S is the number of cascaded logic stages in the critical (longest) path in the digital system. Therefore, system performance is affected critically by both the system design and the performances of the individual of the logic building blocks. For an otherwise similar design, halving the gate propagation delays will allow a doubling of the clock frequency.

Propagation delays are determined experimentally using ring oscillators such as the one shown in Figure 1.23. A ring oscillator comprises an odd number of inverters, so that the voltage at any connection node will oscillate with time. During one period of the oscillation, each gate switches from low to high and then back to low again. Therefore, the period is given by

$$T = \sum_{n=1}^{M} \left(t_{PLH} + t_{PHL} \right), \tag{1.11}$$

where M is the (odd) number of inverters in the ring. If the inverters in the ring are all identical, with identical loading, then the frequency of oscillation is given by

$$f_M = \frac{1}{M\left(t_{PLH} + t_{PHL} \right)} = \frac{1}{2Mt_P}. \tag{1.12}$$

1.2.4 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, and fan-out is the maximum number of load gates that can be connected to the output (without undesirable degradation in performance).

The fan-in may be unity, as in the case of an inverter. However, general system design requires gates with at least two inputs, and gates with higher fan-in may simplify the implementation and improve the overall performance of complex systems. Generally, the maximum practical fan-in is limited by device or circuit performance considerations.

The maximum fan-out N_{MAX} will be determined by loading considerations and is usually estimated using the assumption that the load gates are identical to the driving gates. For MOS-based logic gates, dynamic loading considerations prevail because the loading is primarily capacitive. In this case, 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

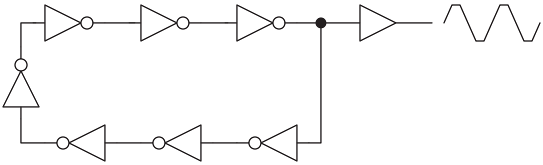


FIGURE 1.23
Seven-stage ring oscillator with a buffered output.

speed. If $t_{P,\max}$ is the maximum allowable propagation delay, then this can be used to determine the maximum allowable load capacitance:

$$C_{L,\max} = f\left(t_{P,\max}\right). \tag{1.13}$$

If the load gates each have an input capacitance C_{in} and other loading effects (self loading, interconnect loading) may be neglected, then the maximum fan-out can be estimated from

$$N_{\max} \leq \frac{C_{L,\max}}{C_{in}}. \tag{1.14}$$

The direct current (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 that flows out of an input lead when logic zero is applied (the input low current), then

$$N_{MAX} < \frac{I_{OL}}{I_{IL}}. \tag{1.15}$$

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 that is sunk by an input with logic one applied, then

$$N_{MAX} < \frac{I_{OH}}{I_{IH}}. \tag{1.16}$$

Typically, the constraints imposed by Equations 1.15 and 1.16 are very different so that the more stringent one prevails.

For CMOS circuits, the dynamic fan-out consideration is prevalent. This leads to an engineering trade-off between fan-out and switching speed. However, system design aspects dictate that the fan-out should generally be greater than 3.

1.2.5 Dissipation

The 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. The cooling of integrated circuits often requires specially designed packages with heat sinking and fans for more efficient heat removal. For VLSI circuits, dissipation considerations may limit the number or density of gates that can be put on a chip.

The dissipation may be dominated by the static component or the dynamic contribution. In some cases, both contributions may be similar in magnitude so that both must be considered.

The dynamic or alternating current (AC) dissipation is associated with charging and discharging of the load capacitance and is called the capacitance switching dissipation. Consider the gate of Figure 1.24 with a lumped capacitive load. 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

$$J = V_{DD} \int_{\substack{clock \\ cycle}} i_{DD} dt \,. \tag{1.17}$$

If there is negligible current flowing from the supply voltage to the ground through the logic gate itself, then during the low-to-high transition at the output, the supply current is given by

$$i_{DD} = C_L \frac{dV_{OUT}}{dt} \,. \tag{1.18}$$

During the high-to-low transition of the output, the load capacitor is discharged through the logic gate but no additional current flows from the supply voltage. Therefore, the energy (in joules) dissipated during one switching cycle is

$$J = V_{DD} \int_{0}^{V_{DD}} C_L dV_{OUT} = C_L V_{DD}^2 \,. \tag{1.19}$$

The dynamic dissipation (in watts) is therefore

$$P_{switch} = fC_L V_{DD}^2 = \alpha f_{CLK} C_L V_{DD}^2 \,, \tag{1.20}$$

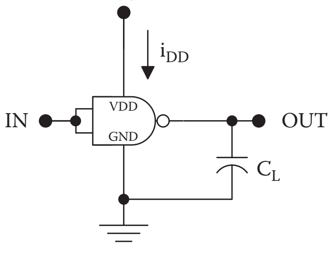


FIGURE 1.24
A logic gate with a lumped capacitive load for the calculation of the dynamic power dissipation.

where f is the switching frequency, f_{CLK} is the clock frequency, and α is the switching activity factor. The activity factor is less than unity, because the actual switching frequency for any particular gate will be less than the system clock frequency. In general, there may be internal nodes of the logic gate, each with its own load capacitance, and these may not all switch at the same frequency as the output node. In that case, the switching power dissipation with K nodes can be estimated from

$$P_{switch} = f_{CLK} \sum_{n=1}^{K} \alpha_n C_{Ln} V_{DD}^2 , \tag{1.21}$$

where α_n and C_{Ln} are the switching activity factor and the load capacitance for the nth node, respectively.

In general, the DC (static) dissipation is dependent on the output state and the fan-out.* In the case of MOS logic gates, the DC load currents are attributable to leakage currents in oxides and p-n junctions as well as subthreshold currents in MOS transistors. Referring to Figure 1.25, the power with a logic zero output state (the "output low power") is

$$P_L = V_{DD} I_{DDL} , \tag{1.22}$$

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.

Referring to Figure 1.26, the power with a logic one output state (the "output high power") is

$$P_H = V_{DD} I_{DDH} , \tag{1.23}$$

where I_{DDH} is the current that flows 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

$$P_{DC} = \frac{P_L + P_H}{2} , \tag{1.24}$$

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

* 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}.

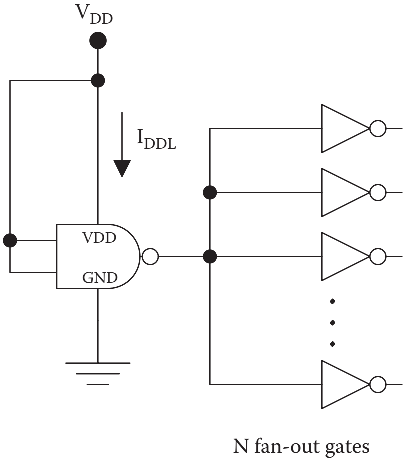


FIGURE 1.25
Determination of P_L (the static output low power).

calculation shown above represents a best estimate and provides a means for the comparison of different logic circuit designs.

The total dissipation is the sum of the DC and dynamic components. Therefore,

$$P = P_{DC} + \alpha f C_L V_{DD}^2 . \tag{1.25}$$

The relative contributions of the DC and dynamic components vary greatly among the different families of logic circuits, but for CMOS circuits, the DC dissipation can nearly always be neglected.

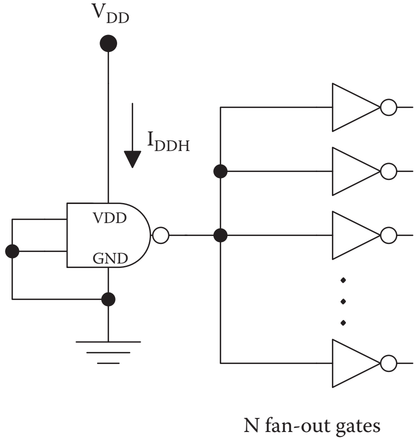


FIGURE 1.26
Determination of P_H (the static output high power).

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

$$PDP = Pt_p, \tag{1.26}$$

where P is the power dissipation per gate, and t_P is the average propagation delay. 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. One interpretation of the power delay product is "the energy required to make a decision." The PDP also serves as an important figure of merit, which relates to the tradeoff between speed and power in a digital logic circuit.

1.3 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 are 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.

Because of the complexity of modern VLSI circuits, computer tools are absolutely necessary to verify designs before fabrication. SPICE (which stands for *simulation program with integrated circuit emphasis*) is a circuit simulation program that was developed at the University of California, Berkeley, in the early 1970s and rapidly became the standard circuit analysis tool in industry. SPICE1 and SPICE2 were developed in the Fortran computer language, and in 1985, SPICE3 was written in C. In the 1980s, Microsim developed PSPICE, a version of SPICE for personal computers. Today, there are several versions of SPICE that run on personal computers or workstations and that go by a number of names, including HSPICE and PSPICE. 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 that are used repeatedly throughout the design. These gates are in turn made with standard transistor designs, and, apart from sizing, there will typically be no more than four distinct types of transistor devices used on a VLSI wafer. Use of standard device and gate designs streamlines the design process greatly, thereby reducing cost and lead time. The standard design languages used for system level design are Verilog and VHDL (which stands for *very high speed integrated circuit hardware description language*).

The device, circuit, and process design are inseparably linked together. Thus, the end product of the design process is a set of computer-generated photomasks. These photomasks are used for pattern transfer to the wafer during the fabrication processing steps.

1.4 Fabrication

The transistors and interconnections in a digital integrated circuit are fabricated on the surface of a semiconductor wafer by a sophisticated series of steps, including epitaxy, oxidation, ion implantation, and thin films deposition (polysilicon, metal, silicon dioxide, silicon nitride). In addition, lithography and etching are used to pattern the semiconductor, insulator, and metal regions as needed to create the transistors and wires on the integrated circuit, using the photomasks referenced in the previous section.

At the present time, most digital logic circuits use the CMOS circuit family, which requires the fabrication of equal numbers of p-MOS and n-MOS transistors on a silicon wafer. Figure 1.27 shows a CMOS inverter circuit and the corresponding physical structure involving one of each type of device, and more complicated logic circuits can be created by the addition of transistors following these two basic designs.

The entire fabrication process starts with a single-crystal wafer of silicon, 300 mm in diameter and 400 μm thick. On this wafer, more than 100 billion transistors and 1000 km of wire will be created during the fabrication of hundreds of integrated circuits. The gate length for the transistors is the most critical dimension, and it is set to the minimum value realizable by the techniques of lithography and pattern transfer; at the present time, manufacturers are transitioning to a printed gate length of 32 nm. Under the gate is a thin (~4 nm) native film of silicon dioxide, created by the oxidation of the silicon wafer. Other thick layers of oxide are created by chemical vapor deposition and serve as insulators between metal and semiconductor or between metal and metal. The doped regions of the semiconductor such as the sources and drains of MOS transistors are created by ion implantation of dopant ions, using patterned oxide to define their placement and dimensions. Metal wires are created in 12 or more levels using copper and aluminum. More than ever before, the device and circuit characteristics are inseparably

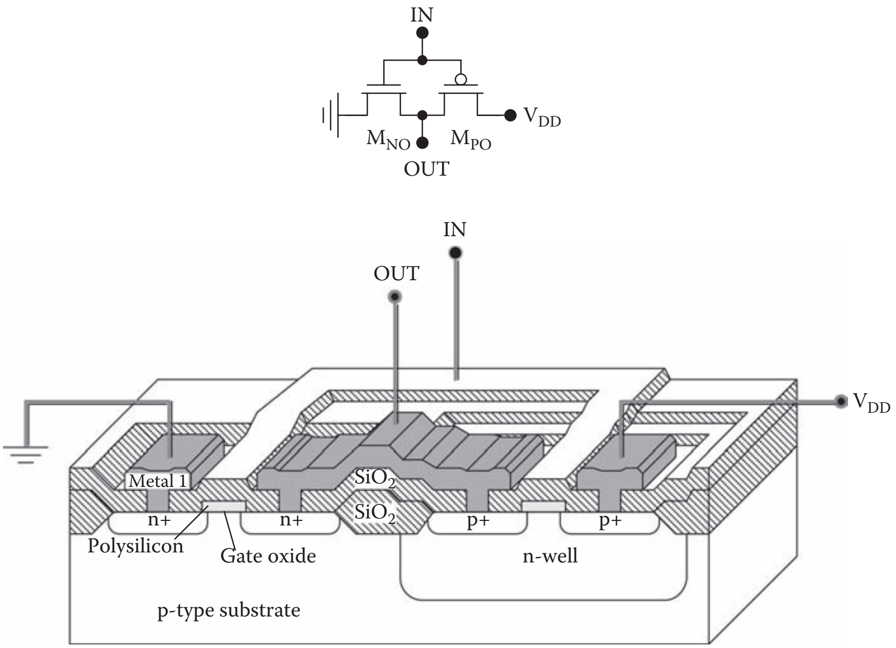


FIGURE 1.27
CMOS inverter (top) and the corresponding physical structure fabricated on a silicon substrate (bottom).

linked to the physical structures and the fabrication process used to obtain them. For example, the vertical junction depths and the horizontal spread of dopants under the gate are critical in determining the device capacitances and therefore delay times. In recognition of this importance, Chapter 2 will describe the fabrication process in greater detail.

1.5 Semiconductors and Junctions

The characteristics of digital circuits are closely linked to the physical properties of the semiconductor material from which they are made. For example, the current versus voltage characteristics of MOS transistors are governed by the carrier mobilities and saturation velocities. All transistors involve p-n junctions in their physical structures, and an understanding of the p-n junction is necessary to fully model the capacitances and leakage currents in MOS transistors. It is in this light that Chapter 3 covers the basic physics of semiconductors and semiconductor junctions. The emphasis is on silicon, the starting material for nearly all digital integrated circuits at the present time and which has the crystal structure shown in Figure 1.28. Other semiconductors, such a silicon germanium alloys and gallium arsenide, have qualitatively similar behavior.

1.6 The MOS Transistor

The MOSFET is the most important device for digital integrated circuits today. It is a unipolar device; that is, electrical current is carried predominantly by the drift of one type of carrier: electrons in the n-MOS transistor and holes in the p-MOS device. This makes the MOS transistor inherently fast switching because minority carrier storage effects are unimportant, and the switching speed is limited predominantly by parasitic capacitances. It is a field effect, or voltage controlled, device that makes the standby power consumption low. These characteristics, and the ease with which the MOSFET geometry can be scaled down in size, make it very attractive for VLSI circuits.

Figure 1.29 shows the basic structure of an n-MOS transistor. The voltage applied at the gate controls the flow of current between the drain and source; when a positive gate-to-source bias is applied, this creates a conducting channel of electrons that move by drift from the source to the drain. In a long-channel MOSFET, carriers drift at a velocity that is proportional to the electric field intensity. In deeply scaled devices, the channel electric field is sufficiently strong so that carriers reach their saturated drift

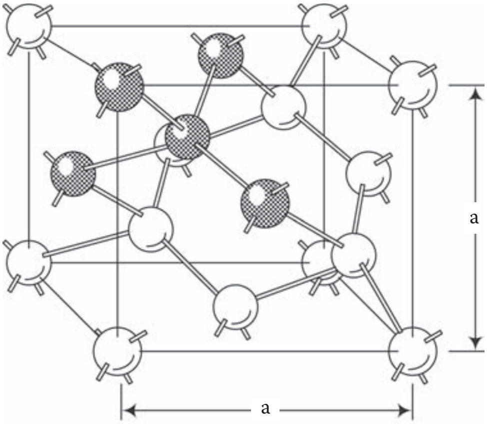


FIGURE 1.28
The diamond crystal structure of silicon. The lattice constant, or the side of the cube, is a = 0.54310 nm. The gate length for a state-of-the-art MOS transistor corresponds to about 60 such cubes lined up in a row.

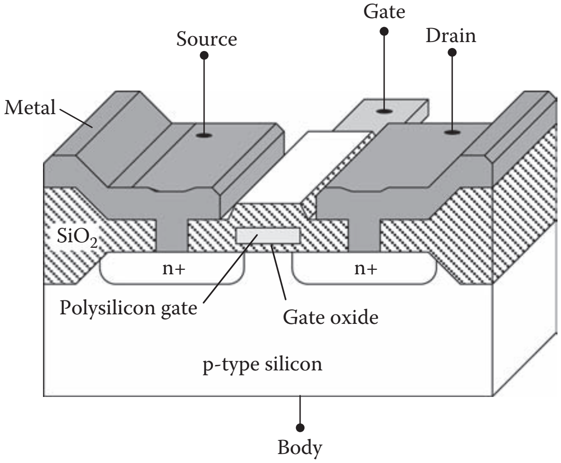


FIGURE 1.29
MOSFET.

velocity (about 9 × 10^6 cm/s for electrons in silicon). This phenomenon, and other effects unique to short-channel MOSFETs, necessitate very complex models for the accurate prediction of digital circuit performance. Chapter 4 describes the physics and SPICE modeling of MOS transistors, starting with long-channel MOS physics and progressing to short-channel phenomena, including the short-channel effect (SCE), the narrow-channel effect (NCE), carrier velocity saturation, and drain-induced barrier lowering (DIBL). There is a detailed description of the SPICE level 1 MOS model, and this model is used for SPICE examples throughout the book. The Berkeley short-channel insulated gate field effect transistor model (BSIM) accurately models the complex behavior of short-channel MOS transistors and is therefore used in advanced VLSI design and industrial practice. Therefore Chapter 4 introduces version one of this model (BSIM1).

1.7 MOS Gate Circuits

*Static combinational logic gate*s are important building blocks in all digital integrated circuits. These circuits realize Boolean functions such as NAND, NOR, and NOT under steady-state conditions using n-MOS transistors (as shown in the example of Figure 1.30) or a combination of complementary n-MOS and p-MOS transistors (as shown in the example of Figure 1.31). Chapters 5 and 6 describe the analysis and design of static combinational gates, starting with the fundamental behavior of inverter circuits (voltage

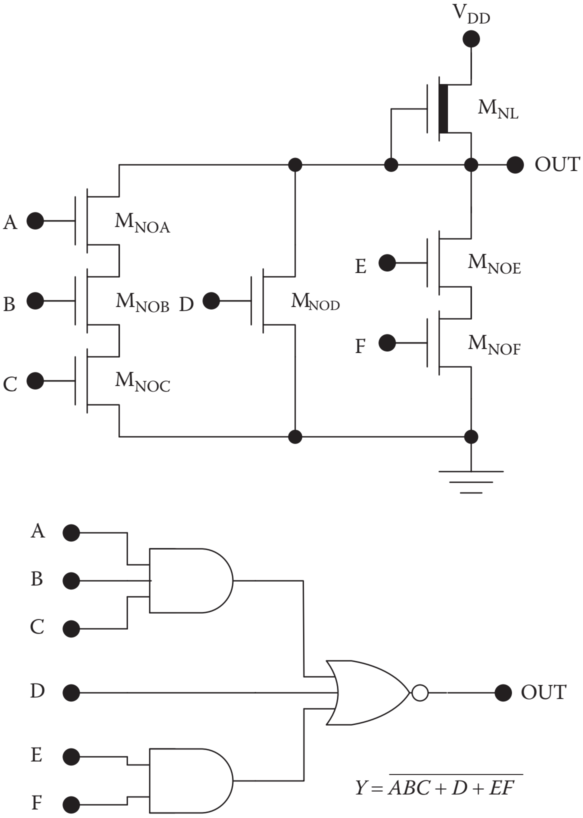


FIGURE 1.30
MOS logic circuit incorporating all n-MOS transistors.

transfer characteristics, delay times, and dissipation). The principles of logic gate design are described, including the sizing of transistors, logic design, and physical layout.

1.8 Interconnect

The wires that connect transistors on a digital integrated circuit, also called *interconnect*, have become increasingly important as transistors have been scaled down in physical size, whereas the number of transistors per chip has doubled every 18–24 months. Integrated circuits now include more than 1 km of interconnect per square centimeter of silicon area, and the parasitic capacitances and resistances of this wiring must be included in delay and

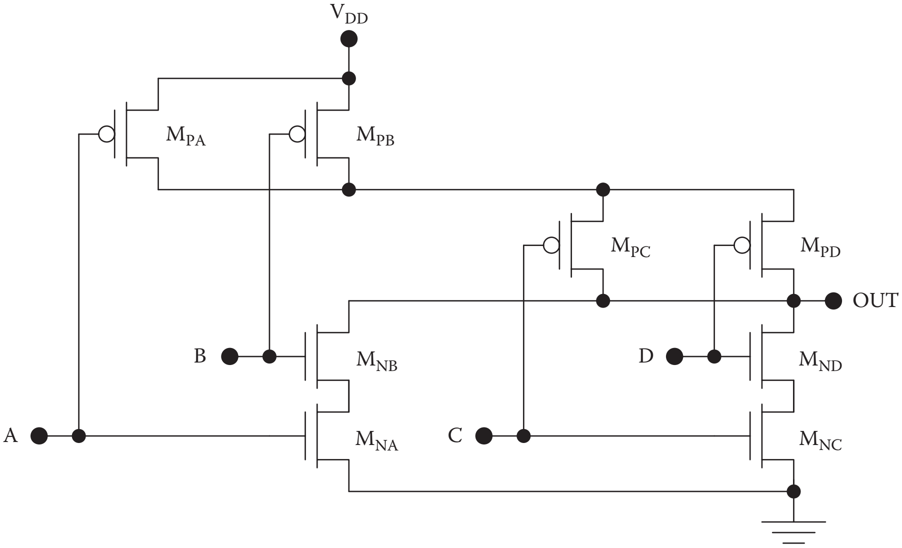


FIGURE 1.31
CMOS AND-OR-INVERT circuit.

power estimates. Chapter 7 describes the materials (copper, aluminum, polysilicon, dielectric) and structures used to realize interconnect and introduces simple electrical models for its behavior in circuits.

1.9 Dynamic CMOS

In addition to static combinational logic gates, which operate under steady-state conditions, integrated circuits often use dynamic circuits that rely on the transient storage of charge on node capacitances. Circuits of this type may achieve higher packing density (logic gates per square centimeter) and lower power dissipation than static gates realizing the same logic functions. Chapter 8 describes simple dynamic CMOS gates, their operation, and characteristics, as well as domino, zipper logic, and dynamic pass transistor circuits.

1.10 Low-Power CMOS

Low-power design has become increasingly important for high-density circuits and battery-operated portable equipment. Whereas battery energy density (in joules per kilogram) has roughly doubled in the past two decades, microprocessor dissipation has increased 50-fold during the same period. Chapter 9 provides a working description of some key approaches to low-power design, including voltage scaling, multiple supply voltages, dynamic voltage scaling, active body biasing, multiple threshold voltages, adiabatic logic, and silicon-on-insulator (SOI). The approaches of active body biasing and SOI underscore the interdisciplinary nature of the digital integrated circuits field and the interplay between materials, device physics, and circuit performance.

1.11 Bistable Circuits

Bistable circuits exhibit two stable states that can represent logic one and logic zero. These include sequential logic circuits such as latches and flip-flops, which are useful in counters, shift registers, and memories. For example, Figure 1.32 depicts a CMOS latch and its two stable states. Bistable circuits can also perform signal shaping functions. An example is the Schmitt trigger, which exhibits hysteresis and is useful in this regard. For example, the CMOS Schmitt trigger of Figure 1.33 with approximately 1 V hysteresis has noise margins superior to those of a simple CMOS inverter.

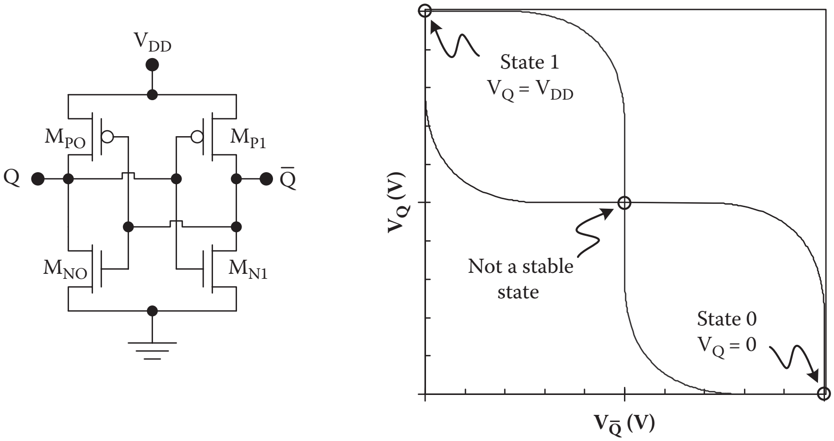


FIGURE 1.32
CMOS bistable latch. The two stable states can be found from the intersections of the characteristics of the two inverters making up the circuit.

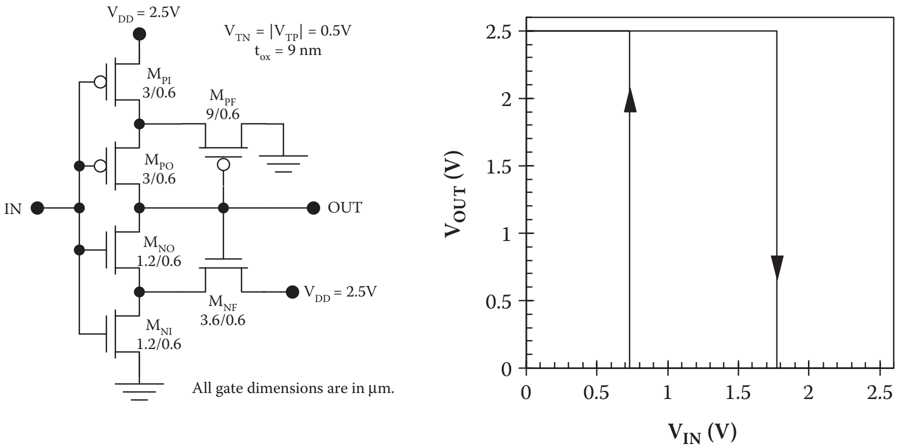


FIGURE 1.33
CMOS Schmitt trigger and characteristic.

1.12 Memories

Memories store code as well as data and are therefore crucial to the operation of any digital processor. Generally speaking, digital memories may be volatile or nonvolatile; nonvolatile memories retain their data when powered down, whereas volatile memories do not. Nonvolatile memories can be further classified as read-only memory (ROM) or read-write memory. Digital memory design is focused on maximizing the functionality (bits); the basis for this is the placement of many identical cells in a rectangular array of rows and columns, as shown in the 4 × 4 NAND ROM of Figure 1.34. Chapter 11 describes the design of basic memory cells, the estimation of memory access delay times, and the basic design of memory row and column decoder circuits.

1.13 Input/Output and Interface Circuits

Digital integrated circuits require the implementation of special input, output, and interface circuits, as well as combinational logic gates, sequential logic, and memory circuits. Inputs must have protection circuitry to prevent electrostatic discharge (ESD) damage during handling as well as transmission gates for enable/disable operation. Output pins generally require high-current output drivers and tri-state operation to allow compatibility with busses. Interface circuits are needed for voltage level shifting between circuits operating with different voltages. Chapter 12 describes these specialized input/output and interface circuits.

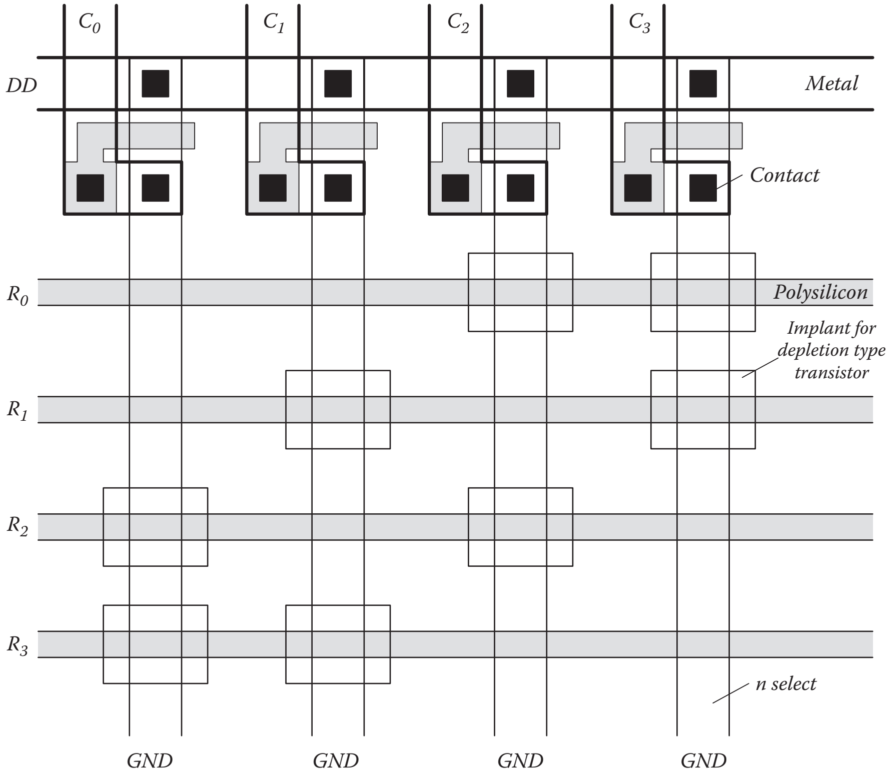


FIGURE 1.34
4 × 4 read only memory constructed with n-MOS transistors.

1.14 Practical Perspective

For practical perspective articles, see the dynamic website at http://www.engr.uconn.edu/ece/books/ayers.

1.15 Summary

Since the invention of the first integrated circuit in 1958, progress in the field of integrated circuits has been extremely rapid: as stated by Moore's law, the number of transistors per chip has doubled every 18–24 months over the past several decades, leading to gigahertz processors and gigabit memories at the present time. A large part of this progress has resulted from the scaling of transistor dimensions, which improves device performance as well as density. As a consequence of this deep scaling of devices, it is now possible to manufacture more than 100 billion transistors and 100 km of wiring on a 300 mm wafer of silicon.

The analysis and design of digital integrated involves the consideration of static combinational logic gates (Chapters 5, 6, and 9), dynamic combinational logic gates (Chapter 8), bistable circuits, including sequential logic gates and Schmitt triggers (Chapter 10), digital memory (Chapter 11), and input/output and interface circuits (Chapter 12). The performance characteristics of modern digital integrated circuits are inseparably linked to the physics of the underlying materials (Chapter 3), devices (Chapters 3 and 4), and interconnects (Chapter 7) as well as the fabrication process (Chapter 2). The complex behavior of MOS devices and circuits necessitate the use of SPICE modeling (Chapters 3–8, 10, and 12) for accurate performance predictions. The BSIM is needed for accurate electrical modeling of deep-submicron MOS transistors and is therefore used in advanced VLSI design and industry.

1.16 Exercises

E1.1. Can you devise a Moore's-type law to describe the growth in dollar sales of integrated circuits for the three decades leading up to 1990, using the data of Figure 1.5? (In other words, how many months did it take, on average, for the dollar sales to double during that period?)

E1.2. If the established historical trends were to continue, in what year would it be possible to produce a flash memory chip (multilevel cell, four bits per cell) holding 1 terabyte (TB)? In what year would a 100 TB chip be produced?

E1.3. An inverter circuit exhibits the voltage transfer characteristic shown in Figure 1.35. Determine V_M, V_{IL}, V_{IH}, and the noise margins.

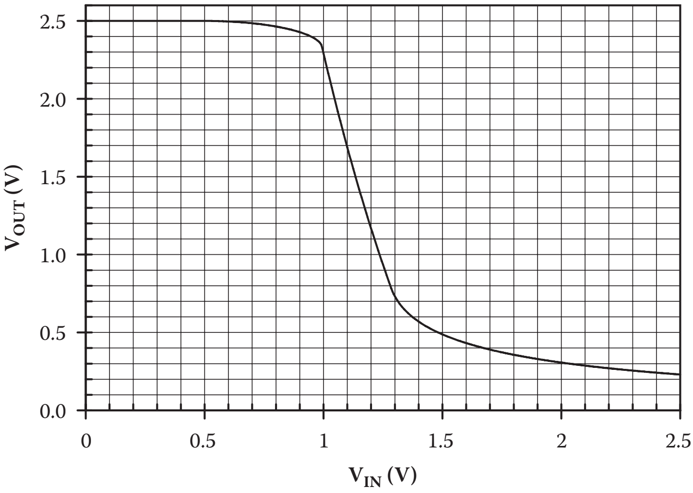


FIGURE 1.35
Inverter voltage transfer characteristic (Exercise E1.3).

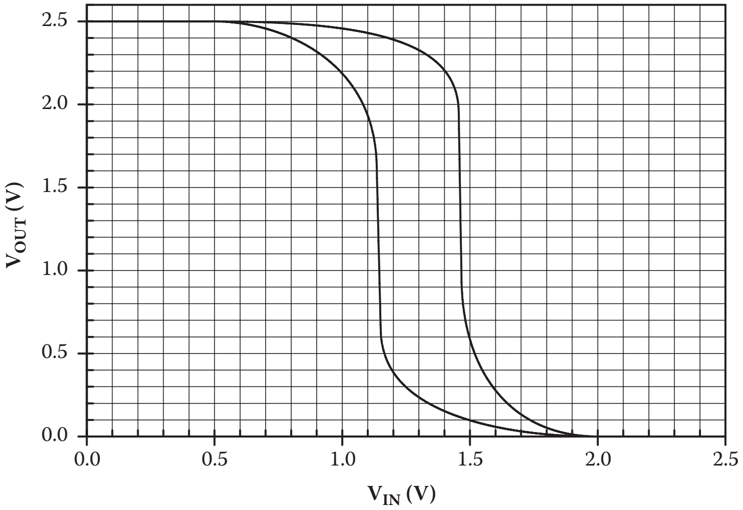


FIGURE 1.36
Static characteristics for a two-way NAND gate (Exercise E1.4).

E1.4. The inputs of a two-way NAND gate exhibit disparate transfer characteristics as shown in Figure 1.36. Determine the worst-case noise margins.

E1.5. From the transient response shown in Figure 1.37, determine the low-to-high and high-to-low propagation delays.

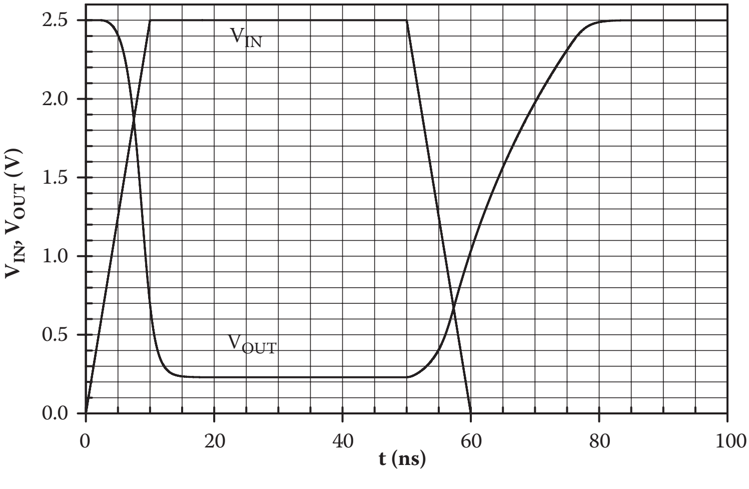


FIGURE 1.37
Transient response for an inverter (Exercise E1.5).

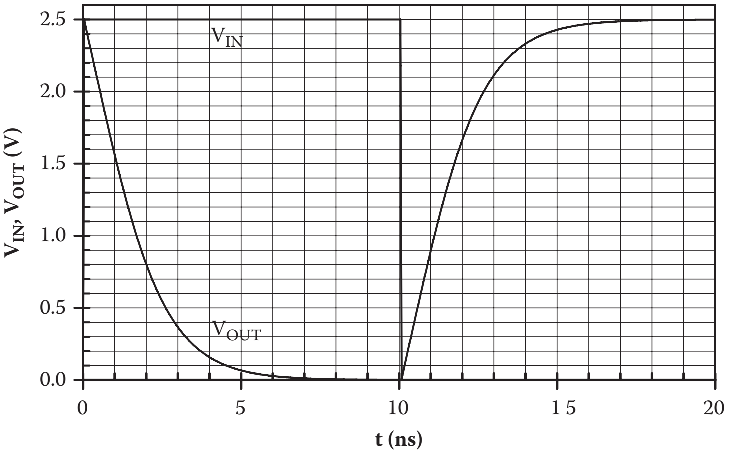


FIGURE 1.38
Transient response of an inverter (Exercise E1.6).

E1.6. From the transient response of Figure 1.38, determine the propagation delays, the rise time for the output, and the fall time for the output. If the rise time is estimated from the 0–90% delay instead of the 10–90% delay, what is the resulting error in the rise time determination? If the fall time is estimated from the 100–10% delay instead of the 90–10% delay, what is the resulting error in the fall time determination?

E1.7. Figure 1.39 shows the output of a five-stage ring oscillator. Each stage is loaded by a 100 fF capacitance. What is the average prop-

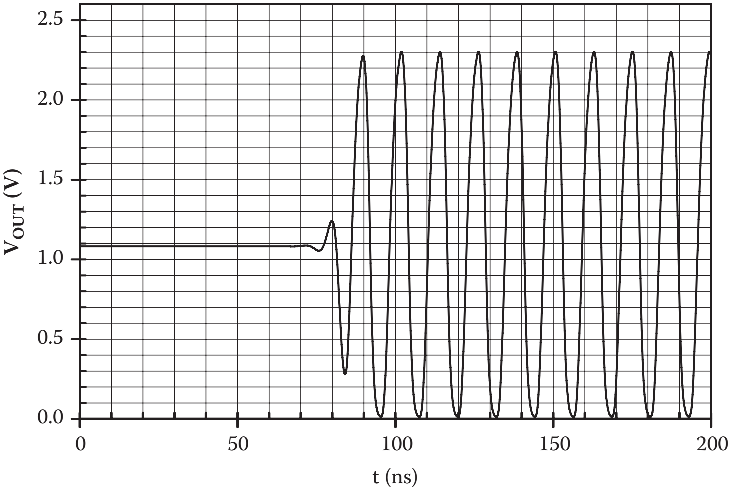


FIGURE 1.39
Output from a five-stage ring oscillator (Exercise E1.7).

agation delay for the gates comprising the ring oscillator? If the supply voltage is 2.5 V, what is the average power consumed by each of the five gates while the ring is oscillating?

For additional exercise problems, see the dynamic website at http://www.engr.uconn.edu/ece/books/ayers.

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 Dev.*, 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 No. 3,025,589 assigned to the Fairchild Camera and Instrument Corp., New York, and U.S. Patent No. 3,029,366 assigned to Sprague Electric Co., Massachusetts.
6. Kilby, J.S., The integrated circuit's early history. *Proc. IEEE*, 88, 109, 2000.
7. Moore, G., Cramming more components into integrated circuits. *Electronics*, 38, 1965.
8. Borkar, S., Obeying Moore's law beyond 0.18 micron [microprocessor design]. *Proc. 13th Annu. IEEE ASIC/SOC Conf.*, 26, 2000.
9. Sinha, A.K., Extending Moore's law through advances in semiconductor manufacturing equipment. *Proc. IEEE First Int. Symp. Quality Electronic Design*, 243, 2000.
10. The Semiconductor Industry Association, San Jose, CA, http://www.sia-online.org.
11. The International Technology Roadmap for Semiconductors, http://public.itrs.net.
12. Lilienfeld, J.E., U.S. Patent 1,745,175, 1930.
13. Kahng, D., and Atalla, M.M., Silicon-silicon dioxide field induced surface devices, IRE Solid-State Device Research Conference, Carnegie Institute of Technology, Pittsburgh, PA, 1960.
14. Arns, R.G., The other transistor: early history of the metal-oxide semiconductor field-effect transistor. *Eng. Sci. Educ. J.*, 7, 233, 1998.

2

Fabrication

2.1 Introduction

The transistors and interconnections in a digital integrated circuit are fabricated on the surface of a semiconductor wafer by a sophisticated series of steps, including epitaxy, oxidation, ion implantation, and thin film deposition (polysilicon, metal, silicon oxide, silicon nitride). In addition, lithographic steps are used to pattern semiconductor, insulator, and metal regions as needed to transfer computer-generated device and circuit designs to the physical wafer. The basic steps of fabrication will be described in this chapter, with examples directed toward the fabrication of planar, complementary MOSFETs. However, it should be emphasized that these same basic processing steps can be adapted to the fabrication of other types of devices, such as dual gate MOSFETs, floating gate MOSFETs, bipolar junction transistors, and even circuits made using other semiconductors.

2.2 Basic CMOS Fabrication Sequence

Figure 2.1 shows the circuit and corresponding physical structure for a CMOS inverter [1–4]. This structure requires one n-MOS transistor and one p-MOS transistor; other logic gates may be realized by the inclusion of additional transistors using these two basic designs. The MOSFETs shown in the figure use a thin (<10 nm) gate oxide and polysilicon gates. There are many variations on the structure shown in Figure 2.1, and it must be recognized that the device and circuit characteristics have become inseparably linked to the physical structure and the fabrication process used to obtain it.

The fabrication sequence starts with large-area, single-crystal wafers of silicon produced by the Czochralski process [5]. Here, a large charge of molten silicon is contained in a graphite or quartz crucible. A nearly perfect seed

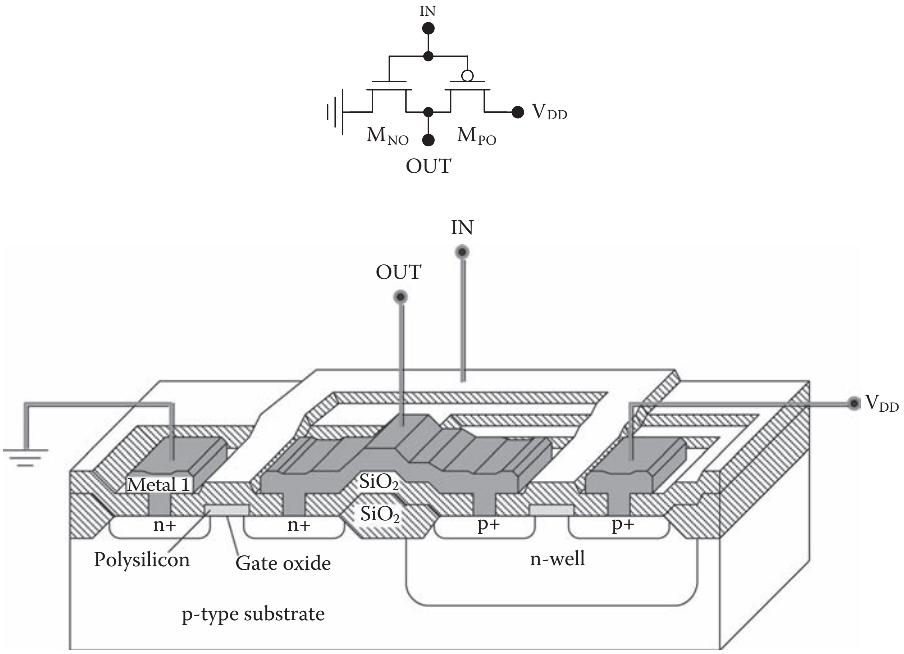


FIGURE 2.1
CMOS inverter (top) and the corresponding physical structure fabricated on a silicon substrate (bottom).

crystal with the desired orientation is brought into contact with the surface of the liquid and then slowly withdrawn while being rotated. In this way, large cylindrical boules of high-quality, single-crystal silicon may be grown. The diameter of the resulting cylindrical boule is determined by the heat input and the pull rate for the crystal, and a feedback control system can be implemented with real-time monitoring of the boule diameter. Moreover, the growing boule can be doped n-type or p-type by the addition of a small quantity of doped silicon powder to the starting silicon charge, and lightly doped p-type wafers are typically used for the fabrication of CMOS integrated circuits. Wafers are produced by slicing the cylindrical boule with a diamond saw, after which these wafers are chem-mechanically polished to a mirror finish.

The first step in the fabrication of CMOS circuits on a p-type silicon wafer is the creation of n-type well regions, in which the p-channel MOSFETs will be placed, as shown in Figure 2.2. This may be accomplished using diffusion or ion implantation. Diffusion involves deposition of a dopant source on the surface of the wafer, which is then subjected to a long heat treatment to cause the solid state diffusion of the dopant atoms into the near-surface region of the wafer. The preferred method of doping is ion

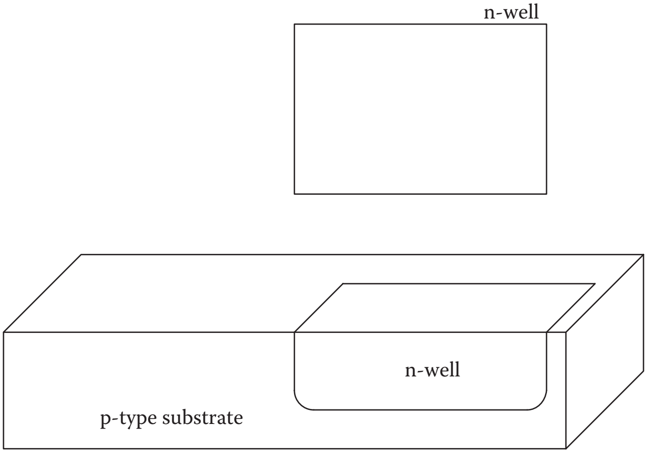


FIGURE 2.2
CMOS process sequence A. Starting with a p-type silicon substrate, an n-well is ion implanted; this well will serve as the *substrate* for the p-channel transistor. The top of the figure shows the mask pattern, and the bottom of the figure shows the resulting physical structure.

implantation, in which dopant ions are accelerated to high energy by a large potential difference (~100–500 kV) and impinge on the semiconductor surface. Most of these ions penetrate the surface and become activated during a subsequent annealing treatment. Advantages of the ion implantation process include precise control of the depth and dose of the dopant, as well as the ability to make extremely shallow junctions required by VLSI circuits.

Whether diffusion or ion implantation is used to create the n-well regions, it is necessary to define the well boundaries using a lithographic step. In this lithographic step, a photosensitive chemical ("photoresist") is applied to the wafer surface and then exposed with ultraviolet radiation using a photomask containing the required pattern. Development of the exposed photoresist results in the transfer of the photomask pattern to the wafer. This patterned photoresist can be used to pattern a secondary mask (such as silicon dioxide [SiO_2]) that is consequently used to establish the boundaries of the doped regions. Similar pattern transfer steps are used many times throughout the wafer fabrication process, with slight variations to accommodate the characteristics of specific materials being patterned.

After the creation of the n-well, the active regions of the transistors are delineated as illustrated in Figure 2.3. Silicon nitride is deposited over the surface area and patterned so that it remains only in the areas defined by the active mask. At the same time, the surrounding silicon is etched below the original surface. This allows the deposition of thick, countersunk SiO_2 surrounding the active regions as shown. This oxide is usually referred to as

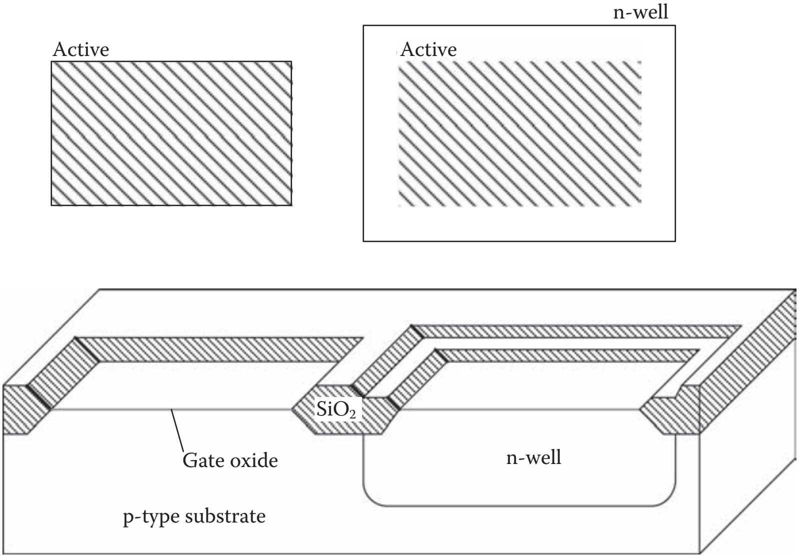


FIGURE 2.3
CMOS process sequence B. "Active" regions are defined by the deposition of silicon nitride. Thick field oxide is produced is all areas not within the active regions using the silicon nitride as a mask. Then the nitride is stripped, and a thin gate oxide is grown over the active regions. The top of the figure shows the mask pattern, with its relationship to the previous mask, and the bottom of the figure shows the resulting physical structure.

"shallow trench isolation." After the nitride is stripped away, a thin (<10 nm) gate oxide is grown by dry thermal oxidation.

After the growth of the gate oxide, polycrystalline silicon (polysilicon) is deposited and patterned with its associated mask as shown in Figure 2.4. Polysilicon is grown by chemical vapor deposition using a source such as SiH_4 and at a temperature sufficiently low that it takes on a fine grain structure. After doping, it will serve as the gates for the MOS transistors. It may also fashion some of the interconnections between devices as shown in the figure.

After the polysilicon has been patterned, the MOSFET source and drain (S/D) regions are doped by ion implantation using the nselect (n-channel transistor) and pselect (p-channel transistor) masks as shown in Figure 2.5. In each case, the polysilicon gate is doped along with the S/D regions, thus diminishing its series resistance. At the same time, the polysilicon gate masks the doping process so that the edges of the S/D regions are automatically aligned with the periphery of the gate. This *self-aligned gate* process has enabled the aggressive scaling of MOSFET dimensions despite tolerances in the alignment of the gate and S/D photomasks.

After doping of the S/D regions, the MOS transistors are ready to be interconnected using metal wires. To this end, an insulating layer such as SiO_2

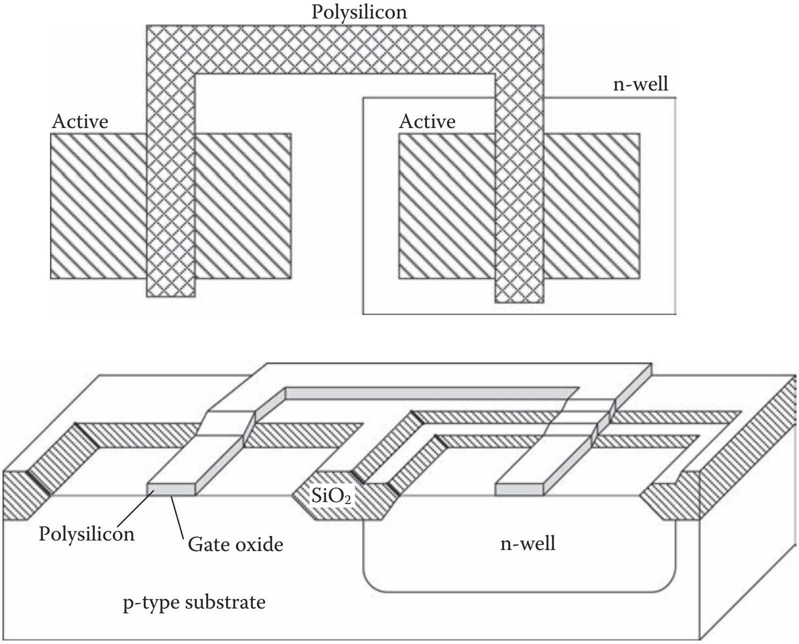


FIGURE 2.4
CMOS process sequence C. Polycrystalline silicon ("polysilicon") is deposited over the entire wafer and then etched in the required pattern to provide gates for the two MOSFETs and also their interconnection. The top of the figure shows the mask pattern, with its relationship to the previous masks, and the bottom of the figure shows the resulting physical structure.

is first deposited over the entire wafer surface and next patterned to allow connection to the underlying devices as shown in Figure 2.6.

On the insulator film, the first metallization layer (metal 1) is deposited and patterned to make connections between the devices as portrayed in Figure 2.7. Aluminum is commonly used for the metal 1 layer for simplicity and low cost. Increasingly, the upper levels of metal are made of copper, whose lower specific resistivity is necessary in high-performance designs. However, the undesirable impurity behavior of copper in silicon necessitates the use of sophisticated barrier layers, and difficulties of etching copper require a novel patterning approach called the damascene method.

Simplified circuit layout diagrams are used in many situations for the purpose of illustration. Figure 2.8 shows the conventional and simplified layout diagram for a CMOS inverter. In a simplified layout diagram such as the one in Figure 2.8b, nselect and pselect areas are omitted and most layer shadings are removed, usually with the exception of the contact areas, which are shown in black, and polysilicon, which is shown in gray. These modifications render the diagram easier to draw and interpret while retaining all information essential for the creation of a detailed layout diagram.

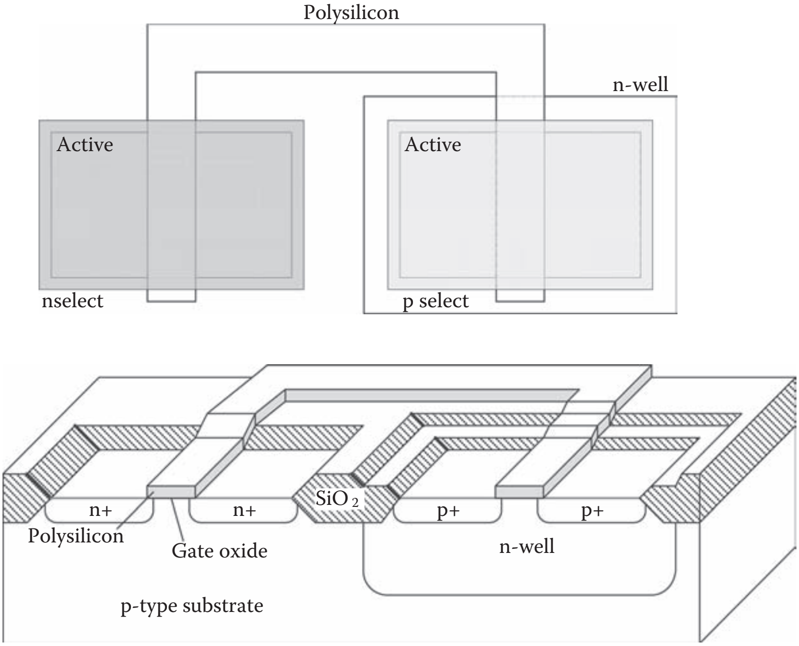


FIGURE 2.5
CMOS process sequence D. Two ion implantation steps are used to dope the S/D regions of the MOSFETs. The n-type doping defined by the nselect mask gives rise to the S/D regions of the n-channel transistor and also dopes the polysilicon gate to reduce its parasitic resistance. The pselect mask allows p-type doping in the corresponding regions of the p-channel device. The top of the figure shows the nselect and pselect mask.

It should be emphasized that modern VLSI circuits use up to 12 or more layers of metal, with intervening insulator layers, thus adding greatly to the number of lithographic steps. This increase in the fabrication complexity, although necessitated by circuit functionality requirements, is accompanied by an increase in cost and a decrease in the yield of good circuits.

2.3 Advanced Processing for High-Performance CMOS

The basic CMOS fabrication process described in the previous section was kept as simple as possible for clarity. Notwithstanding, there are many variations on this process that have been developed for improved performance. These include the use of twin wells, epitaxial layers, and gate spacers. A full description of these and other CMOS process variations is beyond the scope of this book. However, the following subsections will briefly discuss three

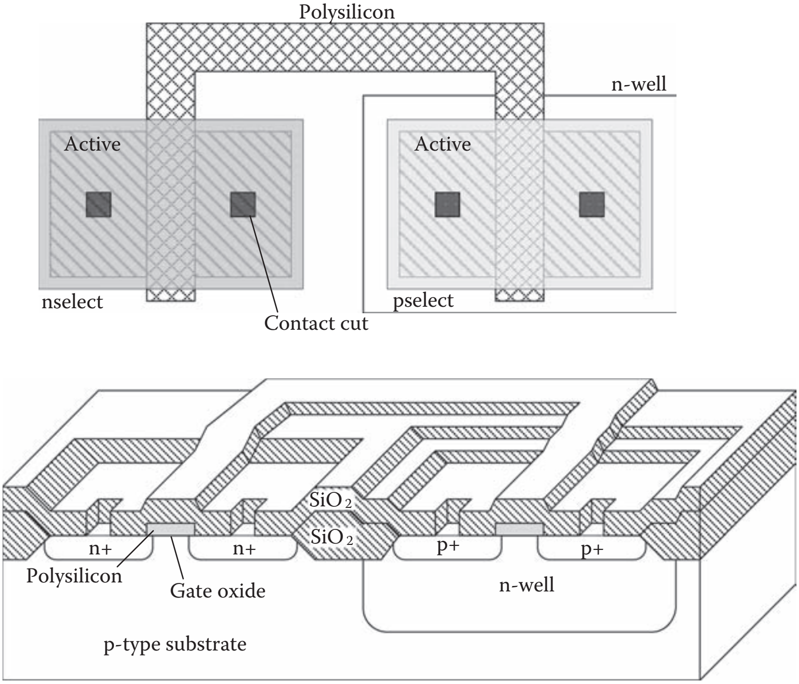


FIGURE 2.6
CMOS process sequence E. After blanket deposition of an insulating layer, contact holes are defined so that metal connections may be made to the devices. The top of the figure shows the mask pattern, with its relationship to the previous masks, and the bottom of the figure shows the resulting physical structure.

key innovations in CMOS technology: these are copper metallization, metal gates, and high-κ dielectrics.

2.3.1 Copper Metal

Copper is superior to aluminum for interconnections between transistors on a VLSI circuit for two reasons: it has a lower specific resistivity, and it is less susceptible to electromigration-induced failure. Despite these advantages, copper was not used until recently because its incorporation in the silicon lattice gives rise to excess leakage currents in devices. However, it has been found that suitable barrier layers, such as tantalum nitride, prevent copper migration into the silicon lattice to eliminate this problem.

The patterning of copper metallization requires a specially adapted approach because the standard technique of reactive ion etching is ineffective with this material. This has led to the development of the dual damascene process, shown schematically in Figures 2.9 through 2.14. These

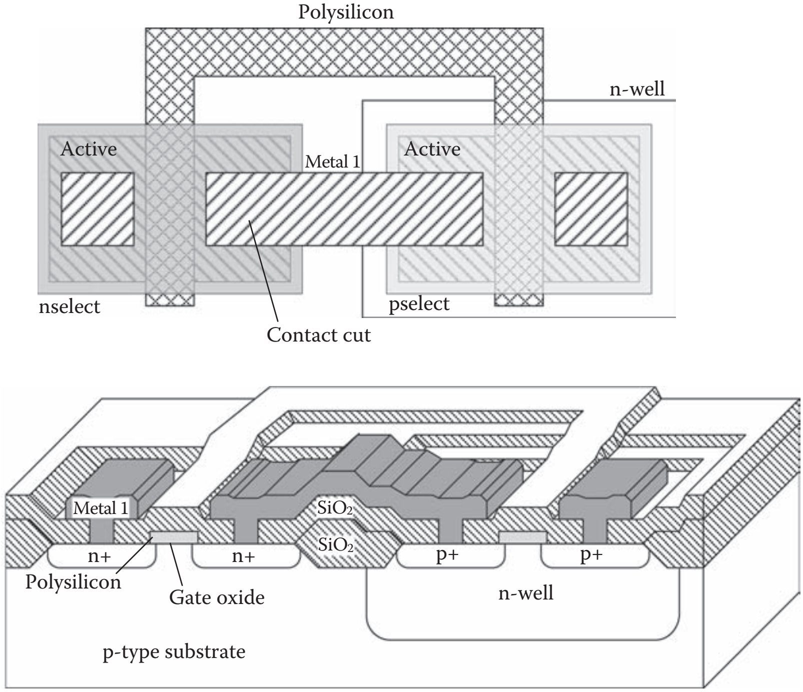


FIGURE 2.7
CMOS process sequence F. The first layer of metallization (metal 1) is deposited and patterned to provide device contacts and interconnections. In a finished circuit, 12 or more layers of metal may be used, with intermediate insulator layers. The top of the figure shows the mask pattern, with its relationship to the previous masks, and the bottom of the figure shows the resulting physical structure.

diagrams portray the use of copper as the level 2 metal on a wafer with the level 1 metal already in place (as shown in Figure 2.9). A three-level stack of oxide-nitride-oxide is deposited as depicted in Figure 2.10. Then a lithographic step is used to etch away the top oxide layer, revealing trenches in which the copper wires will be created. The nitride serves as an "etch stop" by chemically arresting the selective etch process at the depth where it has been placed (Figure 2.11). Next, a *nonselective* etch is used in conjunction with a second lithographic step to create via holes for the connection of level 2 wires with those in level 1 (Figure 2.12). Next, a conformal deposition of copper fills the via holes and trenches and also blanket coats the entire wafer surface (Figure 2.13). Finally, chem-mechanical polishing is used to remove the excess copper, leaving behind only the metal recessed in the via holes and wire trenches; this is depicted in Figure 2.14. The end result is a planar surface, which facilitates the creation of additional levels of metal.

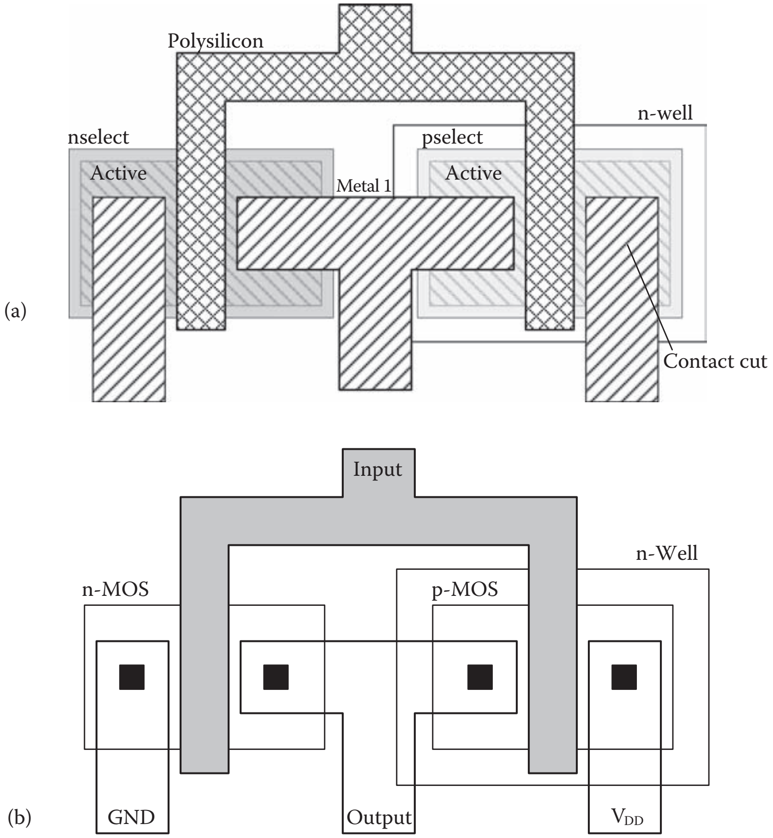


FIGURE 2.8
CMOS inverter: (a) Layout diagram and (b) simplified layout diagram.

The process sequence described here is most commonly called the "dual damascene" process because it involves two lithographic steps to create the wire trenches and via holes before copper deposition. The label "damascene" was applied to this technique because of its resemblance to the ancient metal inlay method.

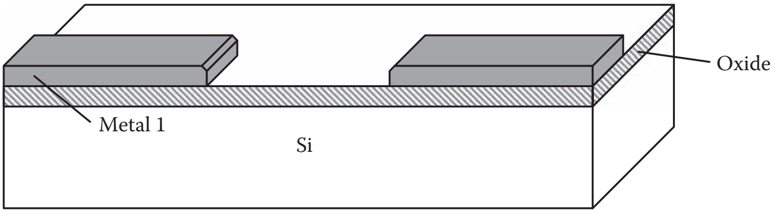


FIGURE 2.9
Copper damascene process sequence A.

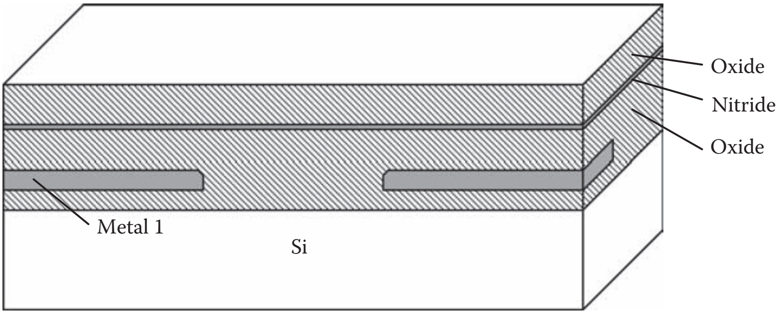


FIGURE 2.10
Copper damascene process sequence B.

2.3.2 Metal Gates

The gate electrode material in modern CMOS circuits is doped polysilicon, which lends itself to the self-aligned gate technology. However, doped polysilicon has a relatively high specific resistivity compared with most metals, and, as a semiconductor, it is susceptible to carrier depletion effects. These two drawbacks have become increasingly problematic as MOSFET devices have been scaled to the nanometer regime. Within the next few generations of VLSI circuits, it will become necessary to replace the polysilicon gate electrodes with metal to alleviate these problems. This switch will introduce new engineering challenges, however. There are few viable materials for this application because the gate electrode material must have appropriate work functions for the p-MOS and n-MOS transistors, and it must also withstand the heat treatment used to anneal the S/D regions unless a sacrificial gate material is used for self-alignment [6]. n-channel and p-channel devices may require *two different metals* to maintain practical work-function differences (see the threshold voltage section in Chapter 4), and this will add one or

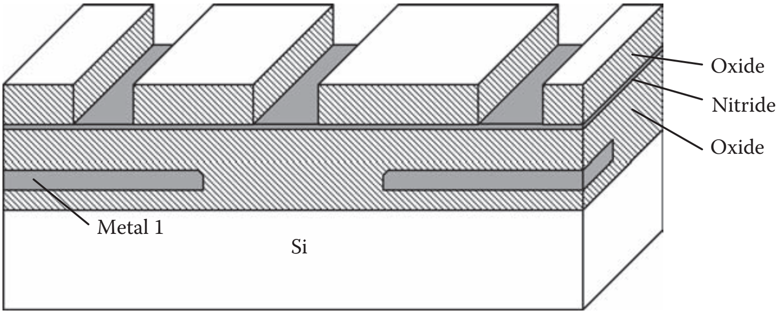


FIGURE 2.11
Copper damascene process sequence C.

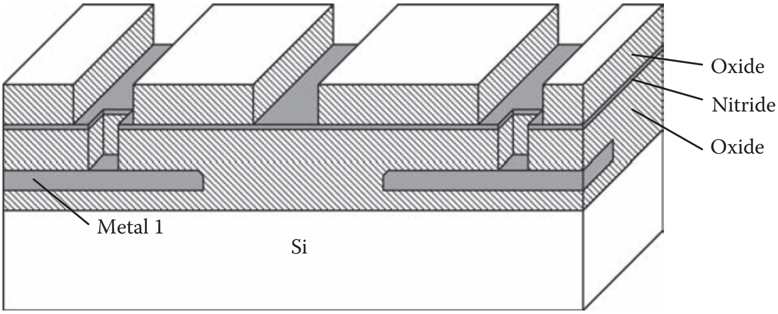


FIGURE 2.12
Copper damascene process sequence D.

more processing steps. One proposed scheme uses TaSiN for the NMOS gate and Ru as the PMOS gate [7]. At the present time, metal gates have already emerged in high-end microprocessors and may soon become the industry standard for high-performance circuits.

2.3.3 High-κ Gate Dielectric

Native SiO_2 has long been used as the gate insulator in silicon MOSFETs because of its high purity and excellent interface properties. However, in modern devices, the oxide thickness has been scaled to the point (~2 nm) at which the tunneling leakage current is becoming intolerable. One solution is the use of a gate insulator having a larger dielectric constant than SiO_2 ("high-κ" dielectric). Such a gate insulator could be made much thicker, ameliorating the excess leakage while maintaining the desired electrical

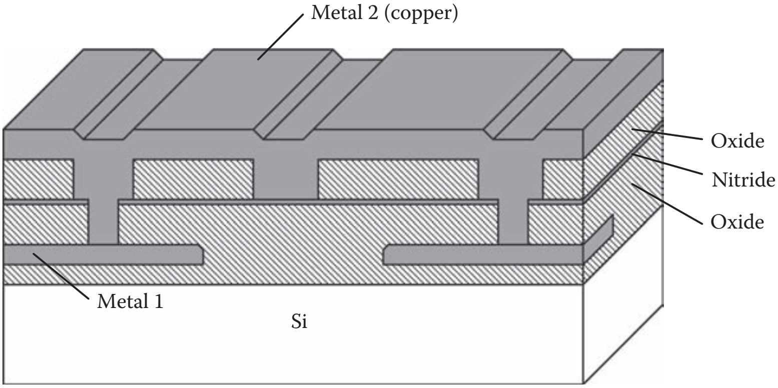


FIGURE 2.13
Copper damascene process sequence E.

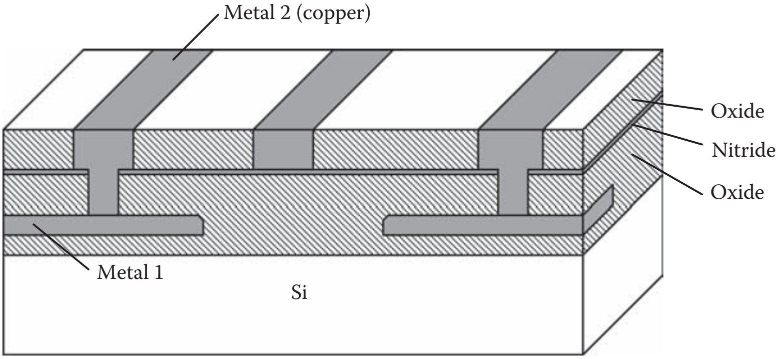


FIGURE 2.14
Copper damascene process sequence F.

behavior of the transistor. Among the available candidates, hafnium dioxide and closely related materials [8, 9] appear best suited for this application and have already made their way to production.

2.4 Lithography and Masks

VLSI fabrication relies on the capability to transfer patterns from computer-aided designs to the physical wafers, by a process called lithography [10–29]. There are several variations on the basic lithographic process, including photolithography [10–16], x-ray lithography [10, 17–24], electron beam (e-beam) lithography [10, 25], ion-beam lithography [26–29], and photoelectron lithography. However, the basis for all of these processes is the exposure and development of radiation-sensitive chemicals called resists [30, 31], which may be positive or negative.

The use of *positive resist* for pattern transfer is illustrated in Figure 2.15. After a fresh layer of SiO_2 is grown over the entire wafer, it is coated with a thin layer of positive resist.* This photoresist is spun on, baked to the desired hardness, and exposed with ultraviolet radiation 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; hence, positive resists are sometimes called *degrading resists*. After development, the patterned photoresist can be used as a mask for the

* Positive photoresists are available from a number of manufacturers, each with their own proprietary formulations. They generally comprise the following: a low-molecular-weight alkali soluble resin (such as phenol formaldehyde novolac), a photoactive dissolution inhibiter (such as orthoquinone diazide), and a solvent (such as xylene).

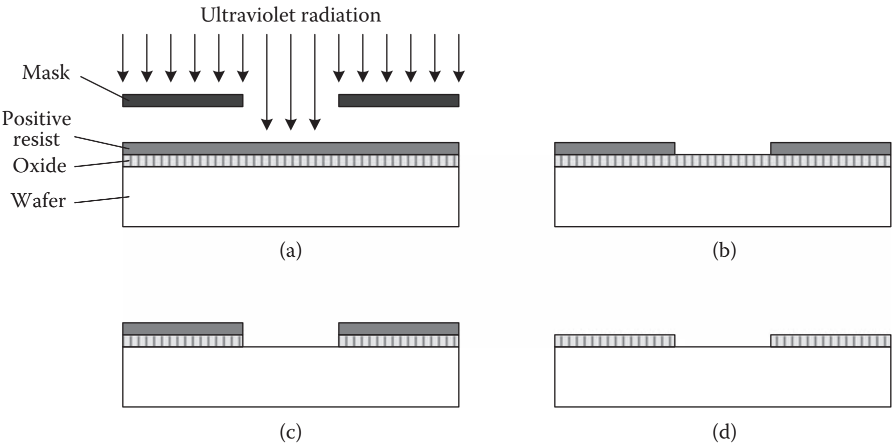


FIGURE 2.15
Pattern transfer using positive photoresist.

chemical etching of the SiO_2. The end result is that the SiO_2 assumes the same pattern as the original mask (a positive image). The SiO_2 can be used as a secondary 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 2.16. 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 that are difficult to remove. For this reason, negative resists are occasionally called *cross-linking resists*. During development, only the unexposed resist is removed, and the remaining resist forms a mask for the patterning of 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 solvent-induced swelling, which occurs during development and 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 patterns are transferred to masks‡ using e-beam lithography, whereby an e-beam is steered directly by the computer using either a raster scan or vector scan approach.

Pattern transfer from the masks to the wafers is done by ultraviolet photolithography using an argon-fluoride excimer laser source because of the

* 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).

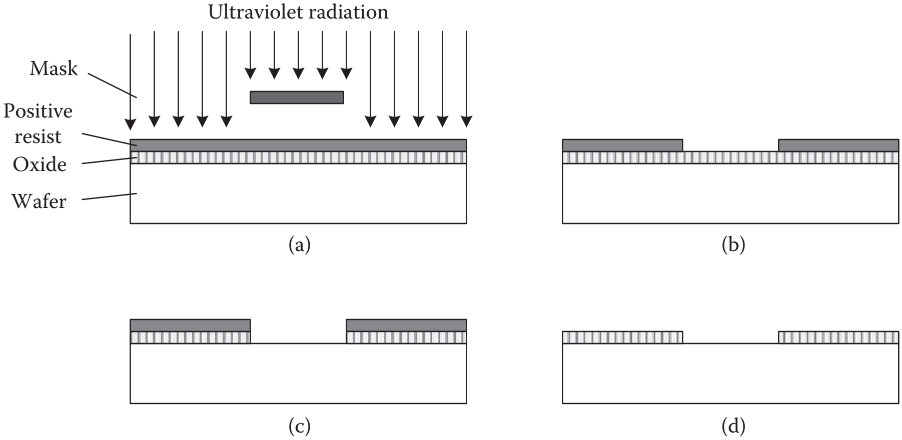


FIGURE 2.16
Pattern transfer using negative photoresist.

higher throughput compared with e-beam lithography. This printing can be done using a contact, proximity, or projection approach as shown in Figure 2.17; however, step-and-repeat projection printing [32–34] is used in all modern VLSI fabrication lines for its extended mask life and excellent resolution compared with the other approaches.

The basic limitations of optical lithography are related to the optical wavelength. The minimum feature size is determined in part by the diffraction limit, given by

$$2X = \frac{\lambda}{NA}, \tag{2.1}$$

where λ is the optical wavelength, and NA is the numerical aperture. The depth of focus for the optical system is

$$d = \frac{\lambda}{(NA)^2}\ . \tag{2.2}$$

For a given optical wavelength, the choice of numerical aperture thus involves a tradeoff between the resolution and depth of focus.

The steady reduction of the minimum feature size from one technology generation to the next has mandated the reduction of the optical wavelength. This is despite the use of phase contrast masks [35] that have allowed resolution performance exceeding the diffraction "limit." At the present time, deep ultraviolet lithography systems are in use and extreme ultraviolet systems [36–38] are being developed for deployment in about 2011.

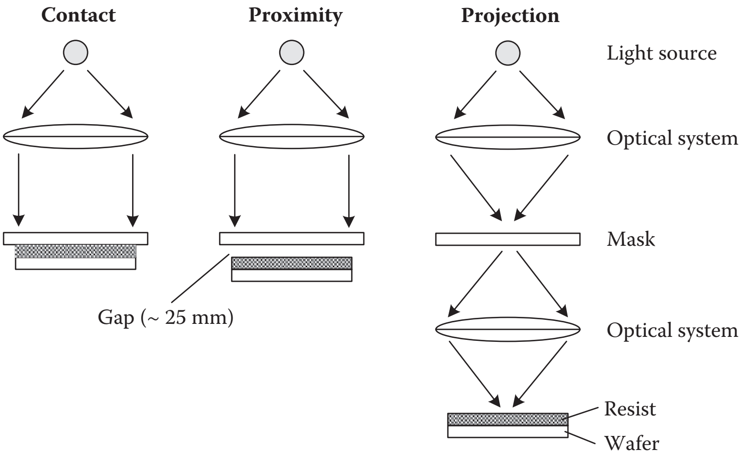


FIGURE 2.17
Contact, proximity, and projection photolithographic printing systems.

2.5 Layout and Design Rules

The physical design of transistors, wires, and other integrated circuit components is governed by a set of design rules. These design rules, which dictate the dimensions, physical size, and device capacitances, are of three types: minimum dimensions, minimum spacings, and minimum surrounds. These minimum values are inextricably tied to the fabrication process as well as the lithographic process used for pattern transfer.

Design rules may be scalable or absolute. Scalable rules are stated in terms of X (where the minimum feature size is 2X)*, whereas absolute design rules are stated in units of length (in nanometers). Scalable rules have the advantage that they can be applied to different process lines having different values of X, but they may not be simultaneously optimized for different values of X. Some design rules do not scale with X, so worst-case values must be used to produce a scalable rule set. In practice, both scalable and absolute design rules are in use today. Examples of scalable design rules sets are those used by the VLSI prototyping service MOSIS [39].

Generally speaking, there are three classes of design rules: (1) minimum widths, (2) minimum spacings, and (3) minimum surrounds. Table 2.1 lists the most important design rules of these types, along with a particular set of values adapted from the MOSIS rule set and used for the examples and exercises in this book. Advanced design work will always be based on similar rules, possibly with the inclusion of additional rules, but the numerical

* 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.

TABLE 2.1

Layout Design Rules along with Values Adopted for This Book

Rule	Description	Value
Minimum dimensions		
L1	Gate length/polysilicon width	2X
L2	Extension of polysilicon gate beyond active region	1X
L3	Width of contact window	2X
L4	Width of active region	3X
L5	Width of implanted region	3X
L6	Width of metal 1	3X
L7	Width of metal 2	3X
Minimum separations		
D1	Spacing between polysilicon gates/wires	2X
D2	Spacing between polysilicon gate and S/D contact window	2X
D3	Spacing between contacts	2X
D4	Spacing between active regions	3X
D5	Spacing between implanted regions of same type	3X
D6	Spacing between metal 1 wires	3X
D7	Spacing between metal 2 wires	4X
D8	Spacing between implanted regions of opposite type	5X
Minimum surrounds		
S1	Active region surrounding contact window	1X
S2	Metal 1 surrounding contact window	1X
S3	Metal 2 surrounding contact window	1X
S4	Polysilicon surrounding contact window	1X
S5	nselect or pselect surrounding contact window	1X
S6	nselect or pselect surrounding active region	2X
S7	n-well surrounding p-MOS active region	5X

values (either scalable, in terms of X, or absolute, in terms of nanomoles) will tend to be different.

These design rules are illustrated in the following series of figures with an orientation to the basic n-well CMOS process described previously. The basic layers and layout legends for such a scalable n-well CMOS process are summarized in Table 2.2.

Note that the active layer defines the placement of silicon nitride, which in turn is used to pattern shallow trench oxide; the shallow trench oxide 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, although these wires exist with both p-type doping and n-type doping because polysilicon is doped simultaneously with the S/D regions of the MOSFETs, as required by the self-aligned process. However, metal or silicide straps are generally placed over the polysilicon to

TABLE 2.2

Legend for Detailed Layouts in a Scalable n-Well CMOS Process

Physical layer	**Name**	**Layout symbol**
n-well	n-well	□
Silicon nitride	Active	▧
Polysilicon	Poly 1	▩
p+ Implant	pselect	░
n+ Implant	nselect	▓
Contact cut	Contact	■
Metal 1	Metal 1	▨
Metal 2	Metal 2	▨

alleviate problems associated with the transition from p-type polysilicon to n-type polysilicon.

Often simplified layouts will be used for the purpose of illustration, and these use the simplified legend of Table 2.3.

2.5.1 Minimum Line Widths and Spacings

The minimum line width 2X is the smallest dimension permitted for any feature in the layout. 2X is also called the *minimum feature size*. At the time of this writing, a minimum feature size of 45 nm is used in production, and we say that we are at the *45 nm technology node*. Technologically, the minimum feature size corresponds to the minimum width for a polysilicon line. For example, with 0.1 μm technology, the minimum polysilicon line width is 0.1 μm, and the value of X is 0.05 μm.

The minimum line widths and spacings are determined primarily by the process technology and equipment used and especially the lithographic process. 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

TABLE 2.3

Legend for Simplified Layout Diagrams

Physical layer	Name	Layout symbol
n-well	n-well	□
Silicon nitride	Active	□
Polysilicon	Poly 1	▒
p+ Implant	pselect	□
n+ Implant	nselect	□
Contact cut	Contact	■
Metal 1	Metal 1	▢
Metal 2	Metal 2	▢

in lateral doping effects. In addition, there are depletion regions surrounding implantations or diffusions made in a semiconductor of opposite conductivity type. Both the lateral doping and the depletion regions affect the minimum spacings of doped regions.

Violation of the minimum line width or spacing rules may result in a nonfunctioning circuit because of broken lines (if the minimum line width is violated) or a short circuit (if the minimum spacing between lines is violated).

Figures 2.18 through 2.22 illustrate the minimum widths and separations for polysilicon, implanted regions, and metal. The minimum width for polysilicon is 2X (rule *L1*) and the minimum polysilicon-polysilicon separation is also 2X (rule *D1*). The minimum width for implanted regions (3X, rule *L5*) 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 (5X, rule *D8*) to reduce the current gain of parasitic bipolar transistors and thereby avoid the latch-up problem. The minimum width for metal 1 is 3X (rule *L6*), and the minimum spacing for metal 1 is also 3X (rule *D6*). The widths and separations for higher levels of metal are generally greater to allow for the loss of planarity on the surface as well as registration errors between mask levels.

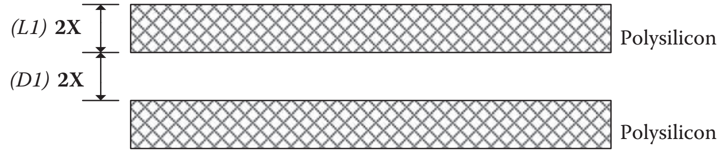


FIGURE 2.18
Polysilicon design rules.

2.5.2 Contacts and Vias

Contacts are made to n+, p+, or polysilicon device regions by opening windows in the overlying oxide before metallization (Figure 2.23). In the scalable rule set adopted here, the minimum dimension for a metal contact is 2X (rule *L3*). 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 contact windows must be spaced by 2X (rule *D3*). The minimum surround for a contact window is 1X for metal 1 (rule *S2*), metal 2 (rule *S3*), polysilicon (rule *S4*), or an nselect or pselect region (rule *S5*). Therefore, the contacting layer as well as the layer being contacted must extend 1X (one-half the minimum feature size) in all directions, allowing for the tolerance in registration between the two mask levels. A special type of contact made between levels of metal is usually referred to as a *via*. An example is the case of a contact made between metal 1 and metal 2.

In Chapter 4, the basic design rules described here will be extended to the design of transistors in several common configurations, and Chapter 6 will illustrate some layout designs for simple CMOS digital circuits.

2.6 Testing and Yield

In the manufacture of VLSI circuits, the yield is of paramount importance and in fact may determine the viability of the manufacturing line or factory.

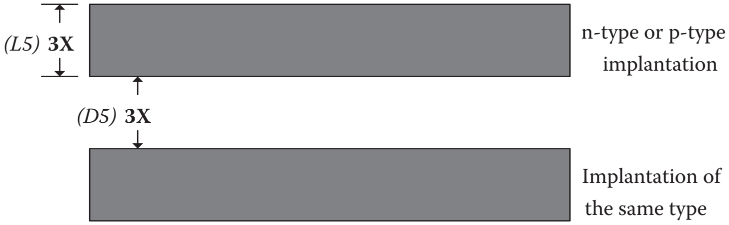


FIGURE 2.19
Design rules for implantations of the same type.

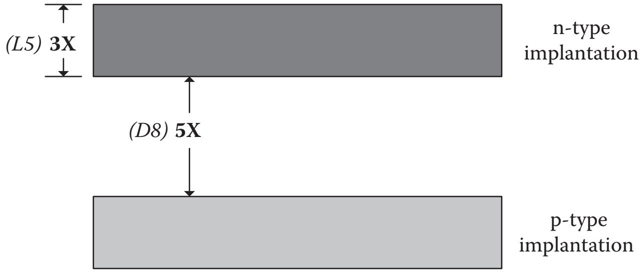


FIGURE 2.20
Design rules for implantations of opposite type.

The yield, which is the fraction of good circuits, is the product of the yields at various steps in the manufacturing:

$$Y = Y_w Y_p Y_a Y_{bi} \tag{2.3}$$

where the wafer yield Y_w is the faction of wafers completing the fabrication process, the process yield Y_p is the fraction of good circuits on finished wafers, before packaging and burn-in, the assembly yield Y_a is the fraction of circuits surviving the packaging process, and the burn-in yield Y_{bi} is the fraction of circuits surviving *burn-in*, which is a stress test at elevated temperature. Typically, the overall yield is determined by the process value, because all of the other factors are close to unity:

$$Y \approx Y_p. \tag{2.4}$$

Wafer testing is always performed on the wafer, before backend processing (dicing, packaging, and burn-in). For this purpose, each circuit is contacted by electrical microprobes that touch the bonding pads or metal bumps on the circuit. 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, thus avoiding the expense of packaging

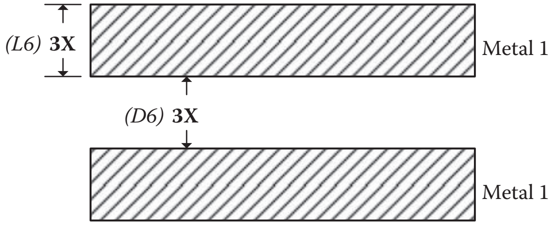


FIGURE 2.21
Metal 1 design rules.

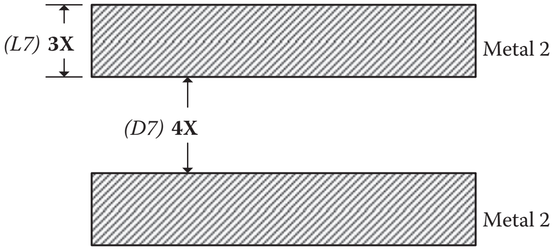


FIGURE 2.22
Metal 2 design rules.

bad die. Generally, both *functional and parametric* tests may be performed. The goal of the functional test (also called the go/no go test) is to determine whether the circuit is functional, whereas a parametric test is performed to evaluate noise margins, propagation delays, and maximum clock frequencies. High throughput is necessary to keep the cost of this step manageable;

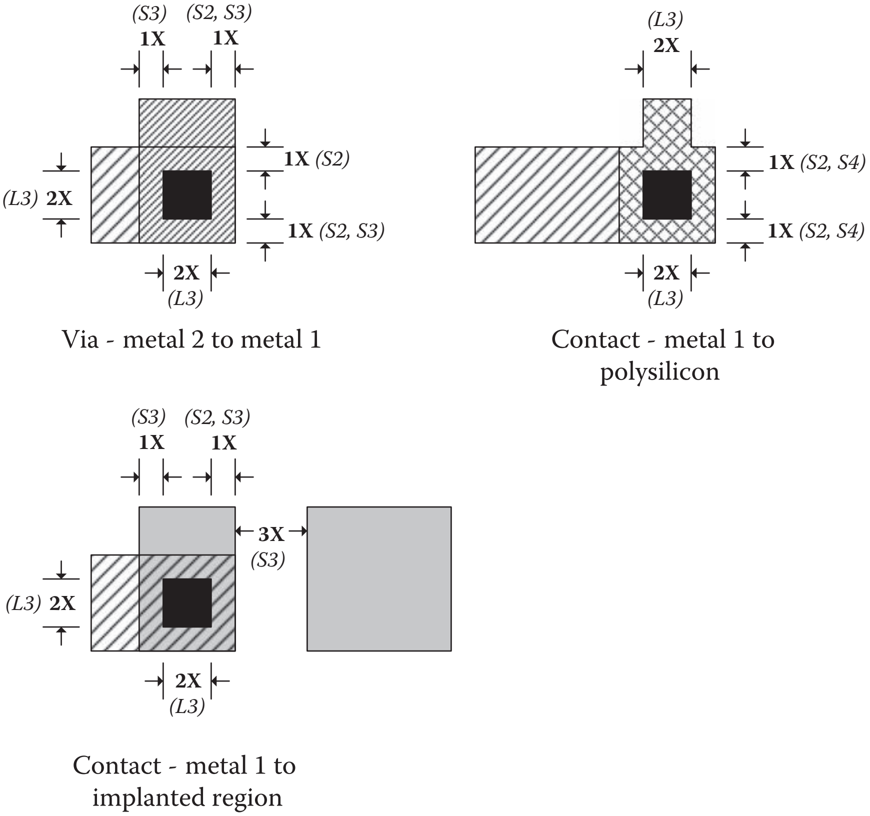


FIGURE 2.23
Layout design rules for contacts.

therefore, testing must be considered at the design phase to enable simple and quick electrical tests.

The processing yield varies greatly from one manufacturing line to another and is also a function of the circuit design being manufactured. Broadly speaking, the underlying causes for bad die may be classified as processing variations and point defects.

Processing variations in doping, oxide thickness, polysilicon 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. Because of these considerations, 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. As a consequence, 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 per cubic meter) greater than 0.5 μm, and class 10 clean rooms have 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 the patterning of the polysilicon and metal 1 layers.

If limited by particulates, the yield is a function of the die size as well as the wafer defect density. Assuming that a single defect produces a nonworking circuit, larger die** will result in a lower yield, even for the same defect density and distribution. This is illustrated in Figure 2.24. The identically sized wafers have the same number and distribution of defects. The wafer on the left has seven bad die of a total of 37. The yield is 30 of 37, or 81%. The wafer on the right with a smaller die size also has seven bad die, but the yield is 81 of 88, or 92%. Whereas a yield of 92% may be economically viable, 81% might not be.

A number of models have been developed for the prediction of the particulate-limited yield. The simplest such model is based on the assumption of a uniform defect density. Suppose that 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 added randomly to the surface of the wafer, the probability of its ruining a good chip is N_G/N. Thus,

$$dN_G = -\left(\frac{N_G}{N}\right) dN_D. \tag{2.5}$$

The fractional yield is therefore

$$Y = \frac{N_G}{N} = e^{-N_D/N} = e^{-D_0 A}, \tag{2.6}$$

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

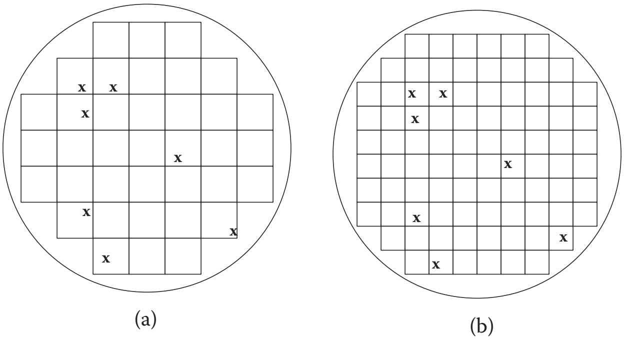


FIGURE 2.24
Effect of die size on yield. Both wafers have the same number and distribution of defects. For wafer (a) the yield is 81%, whereas for wafer (b) the yield is 92%.

where D_0 is the areal density of defects per square centimeter, and A is the die area in cubic centimeters. In practice, this model understates the yield for large area die. For this reason, more complex models have been developed using non-uniform defect distributions.

Integrated circuits of increasing complexity have made it difficult or impossible to completely test the functionality of some chips, such as microprocessors. More and more, circuits are designed for ease of testing. In addition, special test circuits are often built into the wafer for this purpose. Design for test is an important subject covered in books on VLSI design.

2.7 Packaging

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

Any integrated circuit package must meet the following requirements. It must support and protect the integrated circuit from mechanical shock. It should provide protection against the chemical environment, including moisture. It should conduct heat away from the integrated circuit and to an appropriate heat sink. It must be able to withstand a range of operating temperatures and repeated thermal cycling without failure. It must provide the necessary electrical connections to the integrated circuit, without undue degradation of the circuit speed. Finally, all of these requirements must be met in a package that is inexpensive and small in size.

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 use 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 the use of 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 2.25. This package is a type of dual in-line package and has relatively few pins. Other package types can support more than 10^3 pins for VLSI circuits.

Appendix K treats these general principles in more detail. Some important modern package types are also described in this context.

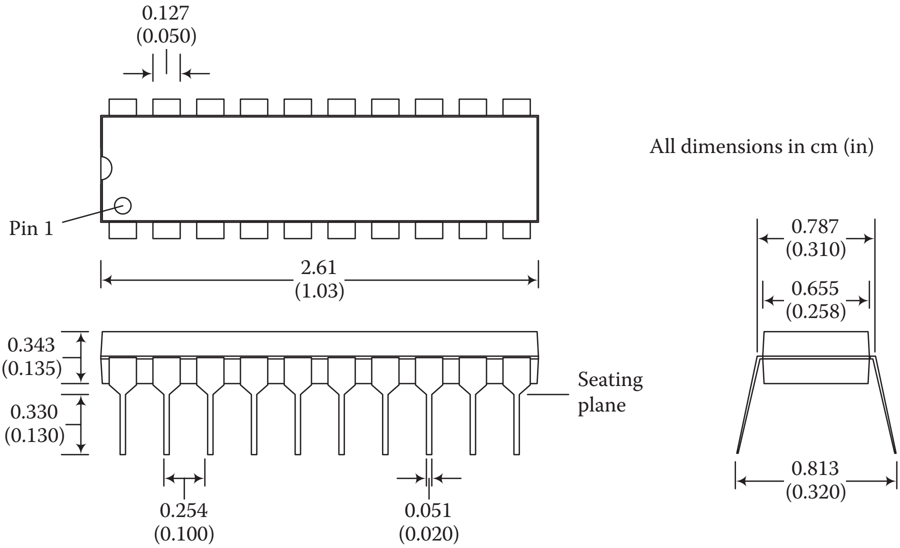


FIGURE 2.25
Plastic dual in-line package (PDIP) with 20 pins.

2.8 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. During burn-in, the integrated circuits are exercised electrically while being 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 so that it does not compromise the reliability of the devices that survive it. The typical failure rates for digital integrated circuits are so low that they may not be measured in a reasonable timescale. 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

$$\text{acceleration} = \exp\left[\frac{E_a}{k}\left(\frac{1}{T_1} - \frac{1}{T_2}\right)\right]. \tag{2.7}$$

For example, with a failure mechanism having 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. An important assumption underlying this analysis is that the failure mechanism is the same at T_1 and T_2; however, this may be impossible to establish if lifetime testing cannot be conducted at the lower temperature.

2.9 Practical Perspective

For practical perspective articles, see the dynamic website at http://www.engr.uconn.edu/ece/books/ayers.

2.10 Summary

Digital integrated circuits are fabricated by a series of steps enabling the creation of more than 100 billion transistors and 100 km of wiring on a single 300 mm wafer of silicon. The overall process involves epitaxial growth, oxidation, ion implantation, and thin film deposition (polysilicon, metal, SiO_2, and silicon nitride), as well as lithographic steps to define the placement and boundaries of doped semiconductor regions, insulators, and conductors. CMOS digital integrated circuits use complementary n-MOS and p-MOS transistors fabricated on the same wafer, as well as 12 or more wiring levels for interconnections. In all, more than 30 lithographic steps may be used to create the finished circuits. Traditionally, integrated MOS transistors have been fabricated with polysilicon gates, native SiO_2 as the gate insulator, and aluminum for contacts and wires. Recent advances in transistor scaling have made it necessary to use metal gates, high-κ dielectrics for the gate insulator, and copper interconnects.

2.11 Exercises

E2.1. The (uniform) areal density of defects for a CMOS fabrication process is 0.2 cm^{-2}. Estimate the yield for integrated circuits with the following dimensions: 4 × 4 mm, 6 × 8 mm, and 9 × 12 mm.

E2.2. When a type of integrated circuit is tested at 175°C, 10% of the circuits fail within 2000 hours with V_{DD} = 1.0 V. Estimate the 10% lifetime at 125°C and 1.0 V assuming E_a = 1.0 eV. Repeat for 100°C.

For additional exercise problems, see the dynamic website at http://www.engr.uconn.edu/ece/books/ayers.

References

1. Davari, B., Koburger, C.W., Furukawa, T., Taur, Y., Noble, W., Megdanis, A., Warnock, J., and Mauer, J., A variable-size shallow trench isolation technology with diffused sidewall doping for submicron CMOS. *IEDM Tech. Dig.*, 92–95, 1988.
2. Davari, B., Koburger, C.W., Schulz, R., Warnock, J.D., Furukawa, T., Jost, M., Taur, Y., Schwittek, W.G., DeBrosse, J.K., Kerbaugh, M.L., and Mauer, J.L., A new planarization technique using a combination of RIE and chemical mechanical polish. *IEDM Tech. Dig.*, 61–64, 1989.
3. Wong, C.Y., Sun, J.Y.-C., Taur, Y., Oh, C.S., Angelucci, R., and Davari, B., Doping of n^+ and p^+ polysilicon in a dual-gate CMOS process. *IEDM Tech. Dig.*, 705–708, 1988.
4. Sun, J.Y.-C., Wong, C.Y., Taur, Y., and Hsu, C., Study of boron penetration through thin oxide with p^+ polysilicon gate. *IEEE VLSI Technology Symp. Tech. Dig.*, 17–18, 1989.
5. Teal, G.K., and Little, J.B., Growth of germanium single crystals. *Phys. Rev.*, 78, 647, 1950.

6. Ren, C., Yu, H.Y., Kang, J.F., Wang, X.P., Ma, H.H.H., Yeo, Y.-C., Chan, D.S.H., Li, M.-F., and Kwong, D.-L., A dual-metal gate integration process for CMOS with sub-1-nm EOT HfO_2 by using HfN replacement gate. *IEEE Electron Dev. Lett.*, 25, 580–582, 2004.
7. Song, S.-C., Zhang, Z., Huffman, C., Sim, J.H., Bae, S.H., Kirsch, P.D., Majhi, P., Choi, R., Moumen, N., Lee, B.H., Highly manufacturable advanced gate-stack technology for sub-45-nm self-aligned gate-first CMOSFETs. *IEEE Trans. Electron Dev.*, 53, 979, 2006.
8. Tsai, W., Ragnarsson, L.-A., Pantisano, L., Chen, P.J., Onsi, B., Schram, T., Cartier, E., Kerber, A., Young, E., Caymax, M., DeGendt, S., and Heyns, M., Performance comparison of sub-1-nm sputtered TiN-HfO_2 nMOS and pMOSFETs. *IEDM Tech. Dig.*, 311–314, 2003.
9. Lee, J.H., Suh, Y.-S., Lazar, H., Jha, R., Gurganus, J., Lin, Y.X., and Misra, V., Compatibility of dual metal gate electrodes with high-κ dielectrics for CMOS. *IEDM Tech. Dig.*, 323–326, 2003.
10. Fukuda, H., and Okazaki, S., Analysis of critical dimension control for optical-, EB-, and x-ray lithography below the 0.2-μm region, Digest of Technical Papers 1995, Symposium on VLSI Technology, 77, 1995. 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 Dev. Mag.*, 19, 43, 2003.
11. Harriott, L.R., Limits of lithography. *Proc. IEEE,* 89, 366, 2001.
12. Brunner, T., Pushing the limits of lithography for IC production, Technical Digest 1997, *Int. Electron Devices Meet.*, 9, 1997.
13. Van den Hove, L., Goethals, A.M., Ronse, K., Van Bavel, M., and Vandenberghe, G., Lithography for sub-90nm applications, Technical Digest 2002, *Int. Electron Devices Meet.,* 3, 2002.
14. Matsuo, T., Endo, M., Kishimura, S., Misaka, A., and Sasago, M., Lithography solution for 65-nm node system LSIs, Digest of Technical Papers 2002, *Symp. VLSI Tech.*, 196, 2002.
15. Pugh, G., Canning, J., and Roman, B., Impact of high resolution lithography on IC mask design**.** *Proc. 1998 IEEE Custom IC Conf.*, 149, 1998.
16. Zacharias, A., X-ray lithography for integrated circuit development and manufacturing. *IEEE Trans. Components Hybrids Manufact. Technol.*, 5, 118, 1982.
17. Murphy, J.B., X-ray lithography sources: A review. *Proc. 1989 Particle Accelerator Conf.*, 2, 757, 1989.
18. Maldonado, J.R., Overview of x-ray lithography at IBM using a compact storage ring. *Conf. Rec. 1991 IEEE Particle Accelerator Conf.*, 542, 1991.
19. 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, Digest of Technical Papers 1998, *Symp. VLSI Tech.*, 82, 1998.
20. Nakayama, Y., Recent progress and future developments in EB mask writing for x-ray lithography, Digest Papers 1999, *Int. Microprocess Nanotechnol. Conf.*, 8, 1999.
21. Uchiyama, S., Current status and issues of x-ray masks. *Proc. 1998 Int. Conf. Microelectronic Test Struct.*, 61, 1998.

22. Mizusawa, N., Uda, K., Tanaka, Y., Ohta, H., and Watanabe, Y., Technology and performance of x-ray stepper for volume production, Digest Papers 2000, *Int. Microprocess. Nanotechnol. Conf.*, 108, 2000.
23. Fukuda, M., and Taguchi, T., Performance of x-ray stepper for next-generation lithography, Digest Papers 1999, *Int. Microprocess. Nanotechnol. Conf.*, 10, 1999.
24. Harriott, L.R., SCALPEL: projection electron beam lithography. *Proc. 1999 Particle Accelerator Conf.*, 595, 1999.
25. Melngailis, J., Focused ion beam lithography and implantation. *Proc. 8th University/Government/Industry Microelectronics Symp.*, 70, 1989.
26. Kim, Y.S., Hong, W., Woo, H.J., Choi, H.W., Kim, K.D., and Lee, S., Ion beam lithography using membrane masks, Digest Papers 2001, *Int. Microprocess. Nanotechnol. Conf.*, 148, 2001.
27. 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 Microelectro. Mechan. Syst.*, 67, 1992.
28. Paek, S.W., Park, S.-H., Lee, H.Y., Chung, H.B., Sub-0.1μm patterning characteristics of inorganic resists by focused-ion-beam lithography. *Proc. Int. Microprocess. Nanotechnol. Conf.*, 129, 1998.
29. Melngailis, J., Ion sources for nanofabrication and high resolution lithography. *Proc. 2001 Particle Accelerator Conf.*, 76, 2001.
30. Rohm and Haas, Philadelphia, PA, http://www.rohmhaas.com.
31. E. I. du Pont de Nemours and Company, Wilmington, DE, http://www.dupont.com.
32. ASML Holding, http://www.asml.com.
33. Canon, http://www.canon.com.
34. Nikon, http://www.nikon.com.
35. Misaka, A., Matsuo, T., and Sasago, M., Super-resolution enhancement method with phase-shifting mask available for random patterns, Digest Technical Papers 2002, *Symp. VLSI Technol.*, 200, 2002.
36. Stulen, R.H., and Sweeney, D.W., Extreme ultraviolet lithography. *IEEE J. Quantum Electronics*, 35, 694, 1999.
37. Gwyn, C.W., Stulen, R.H., Sweeney, D.W., Attwood, D.T., Extreme ultraviolet lithography. *J. Vac. Sci. Technol. B,* 16, 3142, 1998.
38. 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. Microprocess. Nanotechnol. Conf.*, 308, 2001.
39. The MOSIS Service, Marina del Rey, CA, http://www.mosis.com.

3

Semiconductors and p-n Junctions

3.1 Introduction

The active devices in nearly all VLSI circuits are made from the semiconductor silicon. In this single-crystal material, all the valence electrons take part in covalent bonding so the intrinsic conductivity is low. However, the introduction of impurities in trace quantities allows the control of the conductivity and the realization of junctions and transistors, the building blocks of digital integrated circuits. This chapter will review the physical and electrical properties of silicon as well as the physics of p-n junctions.

3.2 Crystal Structure of Silicon

Silicon crystallizes in the diamond structure are shown in Figure 3.1. Like any crystal structure, this can be considered to comprise a space lattice (a periodic arrangement of points in space) and a basis (the arrangement of atoms placed at each lattice point). Here, the space lattice is face-centered cubic with lattice points located at the corners and face centers of the cubic unit cell. The basis includes two identical silicon atoms associated with each lattice point: one on the lattice point and the other displaced by one-quarter of the cube diagonal. In this crystal structure, each silicon atom is bonded covalently to four nearest neighbors located at the apexes of a tetrahedron, as indicated by the shaded atoms in Figure 3.1. The lattice constant, or side of the cubic unit cell, is 0.54311, nm whereas the nearest-neighbor distance is 0.23474 nm [1].

3.3 Energy Bands

The unique properties of semiconductors result from the fact that the allowed energy levels of electrons exist in bands that are separated by

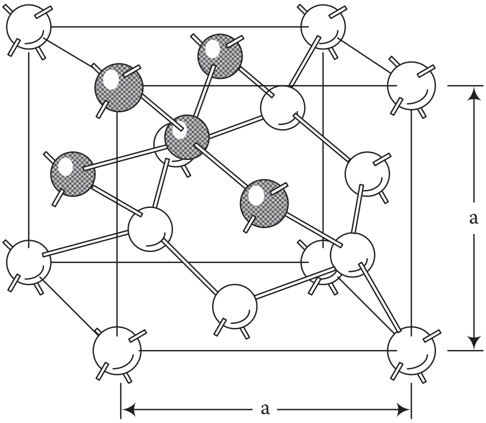


FIGURE 3.1
Crystal structure of silicon. The diamond lattice comprises a face-centered cubic space lattice with a basis of two atoms associated with each lattice point. a = 0.54311 nm.

forbidden gaps. This behavior is attributable to the periodic crystalline structure and is quantum mechanical in nature. For individual atoms making up the crystal, there are discrete allowed energy levels for electrons, but 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 3.2 shows the energy band structure for silicon in an E versus k diagram [2]. (E is the electron energy and k is the electron wave vector. The wave vector directions correspond to the directions in the reciprocal space lattice, and the Γ, L, and X points correspond to points on the first Brillouin zone.) The lower bands are the *valence bands*; in pure silicon at absolute zero, these are completely filled with electrons. The upper bands are the *conduction bands*, which are empty at 0 K in the pure material. An increase in temperature above absolute zero promotes some of the electrons from the top of the valence band to the bottom of the conduction band. The energy required for this transition is the energy gap, 1.12 eV in silicon at room temperature. The presence of electrons in the conduction band and holes in the valence band renders the semiconductor with a small electrical conductivity. In devices, doping with impurities allows the control of the conductivity and type of charge carrier (electrons or holes). For application to devices, it is customary to use simplified energy diagrams in which only the band edges E_C and E_V are shown as functions of distance in one dimension. (For example, see the equilibrium band diagram for a p-n junction in Section 3.9.)

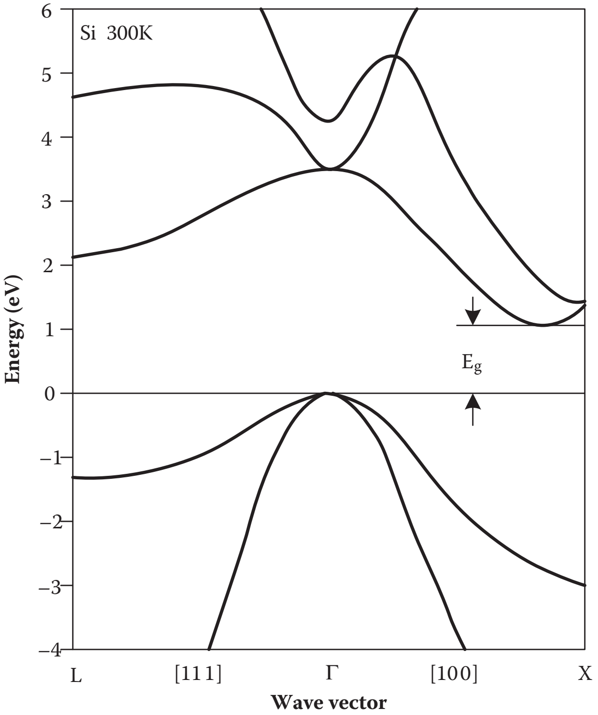


FIGURE 3.2
Energy band structure of silicon at 300 K. (Based on Cohen, M.L., and Chelikowsky, J.R., *Electronic structure and optical properties of semiconductors*, 2nd ed., Springer-Verlag, Berlin, 1988.)

3.4 Carrier Concentrations

With all electrons covalently bonded, there are no free carriers to participate in electrical conduction unless some of the bonds are broken, either thermally, or by illumination with light, or by doping. Regardless of how bond breaking is accomplished, it gives rise to free electrons in the conduction band and free holes* in the valence band, and the concentrations of these free carriers are n and p, respectively.

* Holes are conceptual positive charge carriers used to describe current flow associated with the nearly full valence bands. For a physical description of current transport within such a band, it is more convenient to consider the motion of the holes than the aggregate motion of the remaining electrons. This is analogous to describing the motion of a bubble rather than the surrounding liquid.

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

$$n = N_C e^{-(E_C - E_f)/kT}, \tag{3.1}$$

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

$$p = N_V e^{-(E_f - E_V)/kT}, \tag{3.2}$$

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.

3.4.1 Intrinsic Silicon

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,

$$n = N_C e^{-(E_C - E_f)/kT} = N_V e^{-(E_f - E_V)/kT} = p, \tag{3.3}$$

so that

$$E_f = E_i = \frac{E_C + E_V}{2} - \frac{kT}{2}\ln\left(\frac{N_C}{N_V}\right). \tag{3.4}$$

Both carrier concentrations are equal to the intrinsic value n_i, given by

$$n_i = \sqrt{N_C N_V}\, e^{-E_g/2kT}, \tag{3.5}$$

where E_g is the band gap. At room temperature, E_i is very close to the midgap, and the intrinsic carrier concentration in silicon at room temperature is about 1.45×10^{10} cm^{-3}.

3.4.2 n-Type Silicon

The conductivity of a piece of semiconductor can be altered drastically by doping, which is the controlled introduction of impurities in small concentrations

[*] The Fermi level represents the electron energy at which an existing state has a 50% likelihood of being occupied by an electron. States at lower energies are mostly filled, whereas states at energies higher than E_f are mostly empty.

(usually parts per million). Arsenic, phosphorus, and antimony are pentavalent, with 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 pentavalent atom. It takes very little energy to break this fifth electron away from the impurity and promote it to the conduction band; for this reason, arsenic, phosphorus, and antimony are called electron *donors*.

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,

$$\bar{n} \approx N_d \,, \tag{3.6}$$

where $\bar{n}$ is the equilibrium electron concentration, and N_d is the donor concentration, both per cubic centimeter. In equilibrium, the law of mass action applies so that

$$\overline{np} = n_i^2 \,. \tag{3.7}$$

Consequently, the minority carrier (hole) concentration in this material is given by

$$\bar{p} = \frac{n_i^2}{N_d} \,. \tag{3.8}$$

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).

In very heavily doped material, such as the S/D regions of an n-MOS transistor, the donors will not be fully ionized so that Equation 3.6 no longer holds. This mandates the use of the Fermi integral for a numerical solution of the charge neutrality condition.

Example 3.1 Carrier Concentrations in Doped Silicon

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, so the electron concentration is

$$n \approx N_d = 10^{16} cm^{-3} \,.$$

The hole concentration is

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

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

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

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.

3.4.3 p-Type Silicon

Silicon doped with electron acceptors (usually the trivalent impurity boron) is called p-type. In this material, the acceptor atoms become positively ionized. The carrier concentrations in p-type material are given by

$$\bar{p} \approx N_a \tag{3.9}$$

and

$$\bar{n} = \frac{n_i^2}{N_a} \quad . \tag{3.10}$$

In p-type material, holes are the majority carriers, and electrons are the minority carriers. As in the n-type case, extremely heavily doped material (S/D regions of p-MOS transistors) will exhibit partial ionization of the impurities so that Equation 3.9 will overestimate the hole concentration.

3.5 Current Transport

Electrons and holes are mobile charge carriers capable of carrying electrical current. Underlying this transport is drift and diffusion, in which drift is the motion of carriers in response to an electric field, but diffusion occurs in response to a concentration gradient. In either case, the carriers achieve an average directed velocity that is superimposed on their random thermal motion.

In the one-dimensional case, including both drift and diffusion, the current density equations for electrons and holes, respectively, are

$$J_n = q\mu_n nE + qD_n\frac{dn}{dx} \tag{3.11}$$

and

$$J_p = q\mu_p pE - qD_p \frac{dp}{dx}\,, \tag{3.12}$$

where q is the electronic charge, μ_n and μ_p are the electron and hole mobilities, D_n and D_p are the electron and hole diffusivities, n and p are the electron and hole concentrations, and E is the electric field strength.

For the case of drift alone, the current density is given by

$$J = q\left(\mu_n n + \mu_p p\right)E\,; \tag{3.13}$$

comparison with Ohm's law shows that the conductivity (Ω^{-1} cm^{-1}) is given by

$$\sigma = q\left(\mu_n n + \mu_p p\right) \tag{3.14}$$

and the resistivity (Ωcm) is

$$\rho = \frac{1}{q\left(\mu_n n + \mu_p p\right)}\,. \tag{3.15}$$

The carrier mobilities depend on the temperature and doping concentration. The 300 K mobilities for silicon are given in Figure 3.3 [3].

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

$$D_n = \mu_n \frac{kT}{q} \tag{3.16}$$

and

$$D_p = \mu_p \frac{kT}{q}\,. \tag{3.17}$$

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.

It should be noted that the carrier mobilities provided in Figure 3.3 apply to low-field conditions in bulk material. For the channel region of an MOS transistor, there are additional carrier scattering mechanisms associated with the silicon-insulator interface that reduce the mobility below the bulk values. Typical values are $\mu_n = 580\ cm^2/Vs$ and $\mu_p = 230\ cm^2/Vs$. Also, under

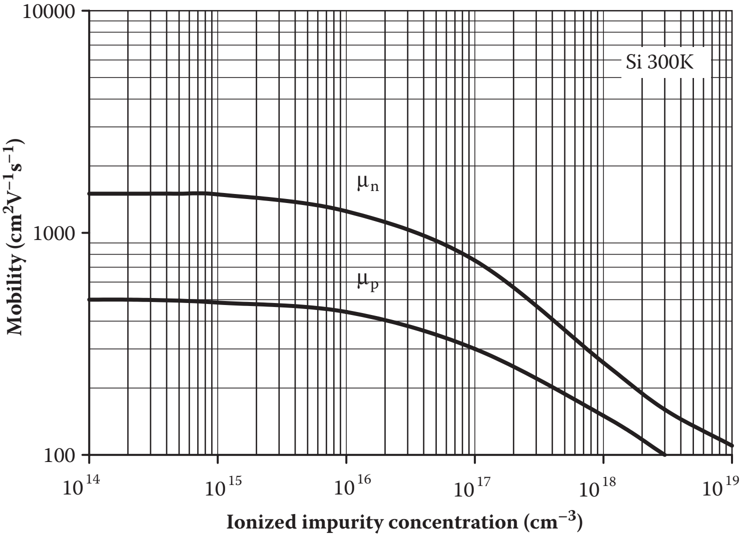


FIGURE 3.3
Electron and hole mobilities in Si versus the ionized impurity (doping) concentration. (Based on Bulucea, C., *Solid-State Electron.*, 36, 489–493, 1993.)

the high-field conditions that exist in short-channel MOSFETs, the carrier velocities tend to reach their saturated values $v_{satn} \approx 9\times10^6 cm/s$ and $v_{satp} \approx 8\times10^6 cm/s$ for electrons and holes, respectively. This behavior can be seen in Figure 3.4, which plots the carrier drift velocities versus the electric field intensity for silicon at room temperature [4, 5].

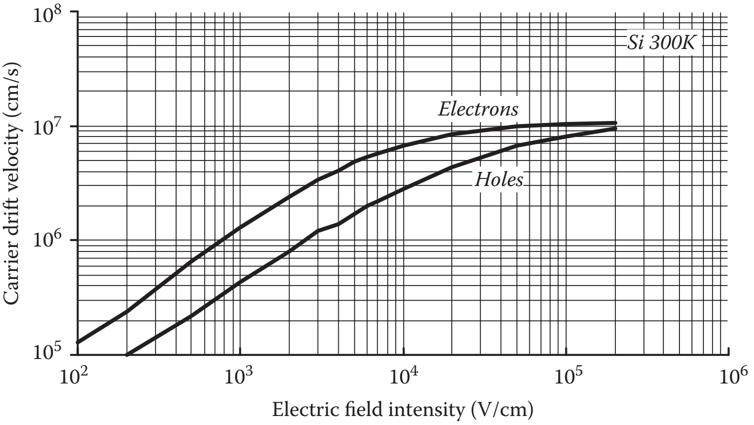


FIGURE 3.4
Carrier velocities versus the electric field intensity for silicon at 300 K. (Based on Jacoboni, C. et al., *Solid-State Electron.*, 20, 77–89, 1977; and Smith, P., Inoue, M., and Frey, J., *Phys. Lett.*, 37, 797–799, 1980.)

3.6 Carrier Continuity Equations

The 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 one-dimensional continuity equation is

$$\frac{\partial n_p}{\partial t} = G - \frac{n_p'}{\tau_n} + \frac{1}{q}\frac{\partial J_n}{\partial x}, \tag{3.18}$$

where n_p' is the excess electron concentration, G is the generation rate for electron-hole pairs, and J_n is the electron current density. Accounting for both drift and diffusion, the current density is

$$J_n = q\mu_n nE + qD_n \frac{dn}{dx}, \tag{3.19}$$

and the one-dimensional continuity equation can be rewritten as

$$\frac{\partial n_p}{\partial t} = G - \frac{n_p'}{\tau_n} + D_n \frac{\partial^2 n_p}{\partial x^2} + \mu_n \frac{\partial}{\partial x}\left(n_p E\right). \tag{3.20}$$

Similarly, for minority carriers (holes) in a n-type semiconductor, the one-dimensional continuity equation is

$$\frac{\partial p_n}{\partial t} = G - \frac{p_n'}{\tau_p} - D_p \frac{\partial^2 p_n}{\partial x^2} + \mu_p \frac{\partial}{\partial x}\left(p_n E\right). \tag{3.21}$$

The continuity equations are the usual starting point for the analysis of devices.

3.7 Poisson's Equation

Another important equation for semiconductor device analysis is Poisson's equation. 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 is

$$-\frac{d^2\psi_i}{dx^2}=\frac{dE}{dx}=\frac{\rho}{\varepsilon}, \tag{3.22}$$

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 attributable to mobile charges (electrons and holes) and fixed charges (ionized acceptors and donors). Hence, Poisson's equation may be written as

$$-\frac{d^2\psi_i}{dx^2}=\frac{dE}{dx}=\frac{q}{\varepsilon}\left(N_d-N_a+p-n\right). \tag{3.23}$$

3.8 The p-n Junction

The fabrication of MOSFETs or other devices in VLSI circuits invariably requires the placement of oppositely doped semiconductors in contact with one another, forming p-n junctions. Important devices in their own right, p-n junctions act as rectifying diodes and can be used for electrostatic discharge protection in VLSI circuits. At the same time, p-n junctions play a role in determining many CMOS device characteristics, such as capacitances, leakage currents, and breakdown voltages. This renders it necessary to review the behavior of p-n junction devices to develop a fuller understanding of VLSI digital circuits.

In a p-n junction, the change from n-type to p-type conductivity results in a transition region, which is depleted of mobile carriers. This *depletion region* takes on a net positive space charge on the n-type side of the junction, where $N_d > N_a$. Similarly, net negative charge exists on the p-type where $N_a > N_d$. This separation of charge gives rise to a capacitance that depends on the bias voltage applied across the p-n junction and must be considered when modeling devices.

Conventional current readily flows in the junction when the p-type region is biased positively with respect to the n-type region (the forward bias condition), because of the injection of holes from the p-region into the n-region and simultaneous injection of electrons from the n-region into the p-region. There is insignificant injection of these carriers to produce a large reverse-bias current, however, so the p-n junction is rectifying and can be used to isolate devices on an integrated circuit. Because the reverse current is not ideally zero, it impacts the standby dissipation in VLSI circuits. Also, the application of a large reverse voltage can cause breakdown, accompanied by a significant reverse current.

The following sections will examine properties of p-n junctions for thermal equilibrium, forward bias, and reverse bias.

3.8.1 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 both 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* that is nearly depleted of free carriers. This region contains net space charge because of the depletion of free carriers. 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 that points 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 3.5. In the quasi-neutral n-type region, the Fermi level E_f is close to the conduction band edge E_C. In the quasi-neutral p-type region, the Fermi level is close to the valence band edge E_V. In the depletion region, there is band bending as described previously. The Fermi level is constant in equilibrium, and this 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

$$\bar{p}_p = N_a \tag{3.24}$$

and

$$\bar{n}_p = \frac{n_i^2}{N_a}. \tag{3.25}$$

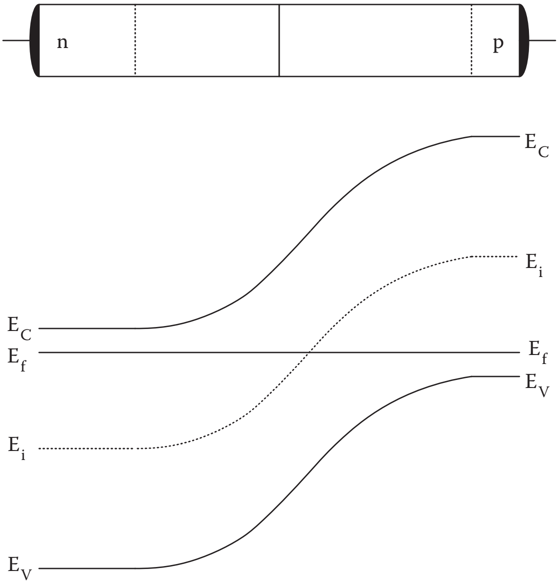


FIGURE 3.5
Equilibrium band diagram for a p-n junction.

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

$$\bar{n}_n = N_d \tag{3.26}$$

and

$$\bar{p}_n = \frac{n_i^2}{N_d}. \tag{3.27}$$

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

$$\rho = -qN_a, \tag{3.28}$$

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 per cubic centimeter. In similar manner, the space charge density on the n-side of the depletion region is

$$\rho = qN_d. \tag{3.29}$$

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

The built-in electric field may be determined using Poisson's equation. Thus,

$$\frac{dE}{dx} = \frac{\rho}{\varepsilon_s}, \tag{3.30}$$

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. It is therefore triangular within the depletion layer and zero outside.

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

$$\frac{dV}{dx} = -E, \tag{3.31}$$

where V is the electric potential. Therefore,

$$V = -\int_0^x E dk. \tag{3.32}$$

This results in a parabolic function of x. The potential difference across the depletion region in thermal equilibrium is called the built-in voltage.

3.8.1.1 Built-In Voltage V_{bi}

The built-in voltage V_{bi} across a p-n junction in thermal equilibrium may be found from the amount of band bending in the junction and is

$$V_{bi} = \frac{kT}{q} \ln\left(\frac{N_a N_d}{n_i^2}\right), \tag{3.33}$$

where 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.

3.8.1.2 Depletion Width W

The width of the depletion region can be determined by combining Equation 3.33 with the condition for overall device charge neutrality, which is

$$x_n N_d = x_p N_a \ . \tag{3.34}$$

Solving, the depletion width on the p-type side is

$$x_p = \sqrt{\frac{2\varepsilon_s V_{bi}}{q}\left(\frac{N_d / N_a}{N_a + N_d}\right)}, \tag{3.35}$$

the depletion width on the n-type side is

$$x_n = \sqrt{\frac{2\varepsilon_s V_{bi}}{q}\left(\frac{N_a / N_d}{N_a + N_d}\right)}, \tag{3.36}$$

and the total depletion width is

$$W = x_p + x_n = \sqrt{\frac{2\varepsilon_s V_{bi}}{q}\left(\frac{1}{N_a} + \frac{1}{N_d}\right)}. \tag{3.37}$$

Figure 3.6 illustrates the space charge density, electric field intensity, and electric potential as functions of distance from the junction, for an abrupt junction in thermal 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

$$W = \sqrt{\frac{2\varepsilon_s V_{bi}}{qN_a}} \text{ (n}^+\text{-p one-sided junction).} \tag{3.38}$$

3.8.2 Depletion Capacitance

The depletion layer exhibits a capacitance that is the same as that of a parallel plate capacitor with a plate separation W. At zero bias, this transition capacitance is

$$C_{JO} = \frac{\varepsilon_s A}{W} = A\sqrt{\frac{q\varepsilon_s}{2V_{bi}}\left(\frac{1}{N_a} + \frac{1}{N_d}\right)^{-1}}. \tag{3.39}$$

Application of a bias voltage modifies the depletion width and therefore the transition capacitance. A forward bias voltage decreases the transition width

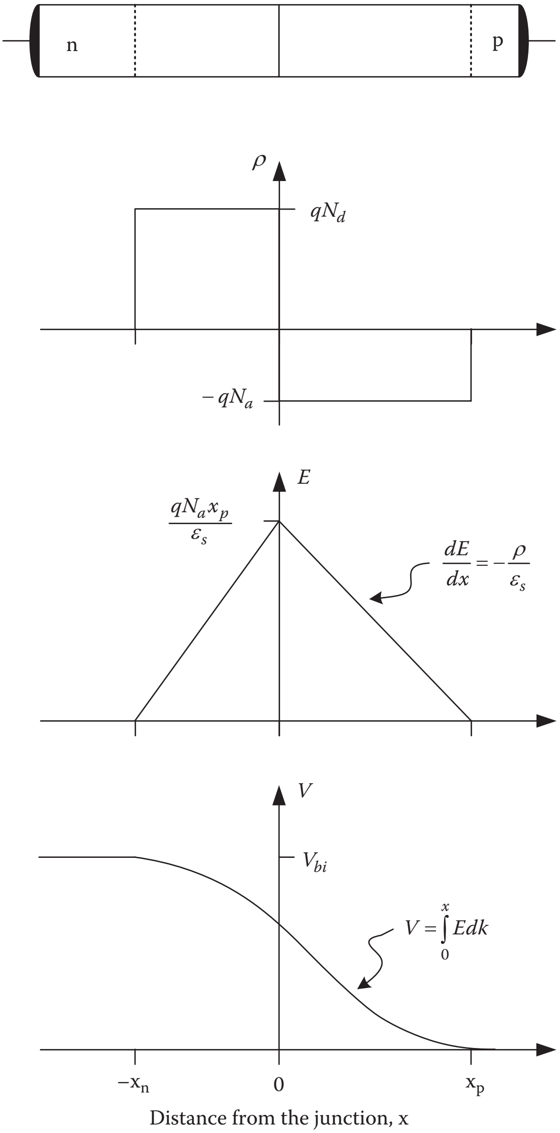

n
p
ρ
qN_d
$-qN_a$
E
$\frac{qN_a x_p}{\varepsilon_s}$
$\frac{dE}{dx} = -\frac{\rho}{\varepsilon_s}$
V
V_{bi}
$V = \int_0^x E dk$
$-x_n$
0
x_p
Distance from the junction, x

FIGURE 3.6
p-n junction in equilibrium (zero bias).

and increases the junction capacitance. Because the modified band bending is $V_{bi} - V$, the bias dependence of the capacitance in an abrupt junction is given by

$$C_J = A\sqrt{\frac{q\varepsilon_s}{2(V_{bi} - V)}\left(\frac{1}{N_a} + \frac{1}{N_d}\right)^{-1}} \tag{3.40}$$

or

$$C_J = \frac{C_{J0}}{\left(1 - {V}/{V_{bi}}\right)^{1/2}}, \tag{3.41}$$

where C_{J0} is the zero-bias value of the capacitance. Here, a positive value of V corresponds to forward bias, but this expression also holds for the case of reverse bias (V < 0).

Junctions with graded doping profiles may be modeled by

$$C_J = \frac{C_{J0}}{\left(1 - {V}/{V_{bi}}\right)^{m}}, \tag{3.42}$$

in which m is the grading coefficient. It has already been seen that $m = 1/2$ is the abrupt junction case; for a linearly graded junction, $m = 1/3$.

Example 3.2 p-n Junction in Equilibrium

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

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

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

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

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

The zero-bias depletion capacitance is

$$C_{J0} = \frac{\varepsilon_s A}{W} = \frac{(11.9)\left(8.85\times10^{-14}F/cm\right)\left(10^{-5}cm^2\right)}{0.33\times10^{-4}cm}$$

$$= 0.32\times10^{-12}F = 0.32pF\ \cdot$$

3.8.3 Forward Bias Current

Under forward bias conditions (p-region biased positively with respect to the n-region), the magnitude of the current increases exponentially with the applied bias [6–9]. In developing a model for the forward current, the usual approach is to consider forward bias as a small departure from thermal equilibrium as described in what follows.

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

$$\bar{n}_p = \bar{n}_n \exp\left(-\frac{qV_{bi}}{kT}\right), \tag{3.43}$$

where $\bar{n}_p$ is the equilibrium concentration of electrons (minority carriers) on the p-type side, and $\bar{n}_n$ 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

$$n_p(x=0) = \bar{n}_n \exp\left(-\frac{q(V_{bi}-V)}{kT}\right), \tag{3.44}$$

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,

$$n_p(x=0) = \bar{n}_p \exp\left(\frac{qV}{kT}\right), \tag{3.45}$$

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, and 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 junction 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

$$\frac{\partial n_p}{\partial t} = -\frac{n'_p}{\tau_n} + D_n \frac{\partial^2 n_p}{\partial x^2}, \tag{3.46}$$

* The *base* of a p-n junction is the more lightly doped region. Forward current conduction is dominated by the injection of minority carriers into the base.

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

$$n_p' = n_p - \overline{n}_p . \tag{3.47}$$

The boundary conditions are

$$n_p'(x=0) = \overline{n}_p \left(e^{qV/kT} - 1\right) \tag{3.48}$$

at the edge of the depletion layer and

$$n_p'(x = W_B) = 0 \tag{3.23}$$

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

$$n_p'(x) = \overline{n}_p \left(e^{qV/kT} - 1\right) \frac{\sinh\left(\dfrac{W_B - x}{L_n}\right)}{\sinh\left(\dfrac{W_B}{x}\right)} , \tag{3.49}$$

where $L_n = \sqrt{D_n \tau_n}$ is called the *minority carrier diffusion length*. This behavior is shown in Figure 3.7.

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

$$J_n\left(x=0\right) = qD_n \left.\frac{\partial n_p'}{\partial x}\right|_{x=0} = \frac{qD_n \overline{n}_p}{L_n \tanh\left(W_B / L_n\right)} \left(e^{qV/kT} - 1\right). \tag{3.50}$$

Multiplying by the junction area, we obtain the current

$$I_n = \frac{qAD_n \overline{n}_p}{L_n \tanh\left(W_B / L_n\right)} \left(e^{qV/kT} - 1\right) = \frac{qAD_n n_i^2}{L_n N_a \tanh\left(W_B / L_n\right)} \left(e^{qV/kT} - 1\right). \tag{3.51}$$

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

$$I = I_S \left(e^{qV/kT} - 1\right), \tag{3.52}$$

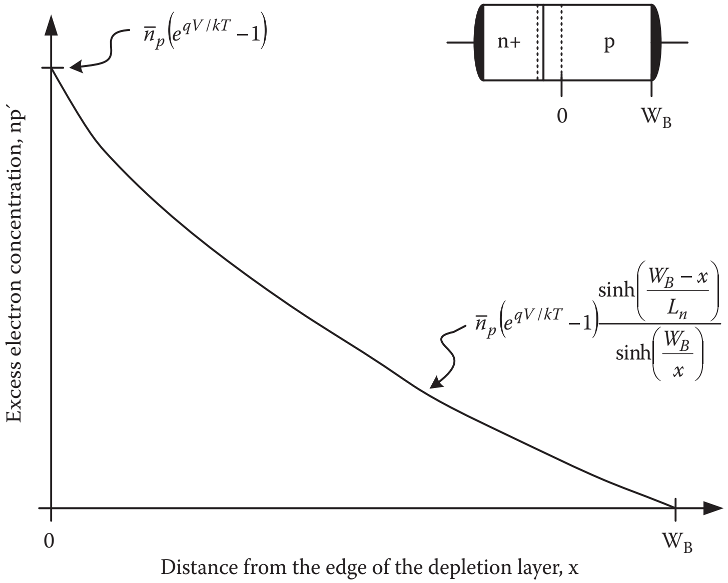


FIGURE 3.7
The excess minority carrier (electron) concentration versus distance from the edge of the depletion region for an n+-p junction under forward bias conditions.

where the reverse saturation current I_S for an n+-p junction is given by

$$I_S = \frac{qAD_n n_i^2}{L_n N_a \tanh\left(W_B / L_n\right)}. \tag{3.53}$$

The diode equation can be simplified in certain practical cases as shown in the following subsections [10].

3.8.3.1 Short-Base n+-p Junction

A short-base n+-p junction is one in which the width of the p-type base is much less than the minority carrier diffusion length in that region. Thus, $W_B << L_n$ so that the excess minority carrier profile is linear as shown in Figure 3.8.

The resulting diffusion current is therefore

$$I_n \approx \frac{qAD_n n_i^2}{N_a W_B}\left(e^{qV/kT}-1\right). \tag{3.54}$$

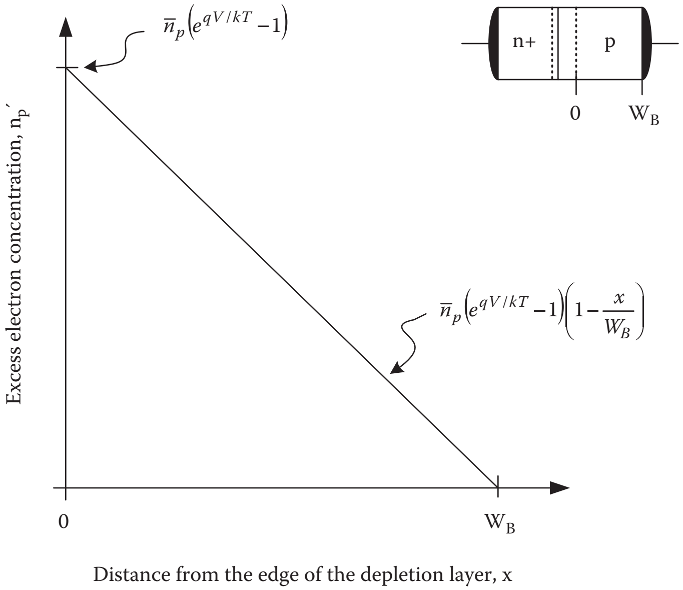


FIGURE 3.8
Excess minority carrier (electron) concentration as a function of distance from the edge of the depletion region for a short-bias n+-p junction with forward bias.

Example 3.3 Forward Biased p-n Junction

Consider an n+-p junction at 300 K with an area of $2\times10^{-6}cm^{-2}$. For the n+ emitter, $N_d = 10^{19}cm^{-3}$, $D_p = 2.0\,cm^2s^{-1}$, and $\tau_p = 1ns$. For the p-type base, $N_a = 10^{16}cm^{-3}$, $D_n = 15\,cm^2s^{-1}$, and $\tau_n = 5ns$. 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 junction. The diffusion length for electrons in the base is

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

Therefore, $L_n >> W_B$, and this is a short-base junction. The saturation current is

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

$$= 5.0\times10^{-15}A\,\cdot$$

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

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

3.8.3.2 Long-Base n+-p Junction

A long-base n+-p junction is one in which the width of the p-type base is much greater than the minority carrier diffusion length in that region, or $W >> L_n$ so that

$$I_n \approx \frac{qAD_n n_i^2}{N_a L_n}\left(e^{qV/kT} - 1\right). \tag{3.55}$$

3.8.4 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 junctions with moderate reverse bias. With larger reverse bias voltages applied, there is even the possibility of reverse breakdown, in which the current is limited only by the external circuit.

With 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], and this 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 so that 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

$$I_{gen} = -\frac{qAWn_i}{\tau_{sc}}, \tag{3.56}$$

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 attributable to the bias dependence of the depletion width (square root dependence).

The space charge lifetime is related to, although not equal to, the minority carrier lifetimes in the quasi-neutral p-type and n-type regions. However, the values of lifetime are all shortened by the introduction (intentional or unintentional) of *lifetime killer* impurities such as gold, nickel, or platinum. High-frequency silicon 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 increase reverse leakage currents and standby power dissipation without providing any benefits in switching speed.

3.8.5 Reverse Breakdown

Under the condition of a sufficiently large reverse bias, junction breakdown occurs; this is accompanied by a large reverse current, often limited only by the external circuit. This condition is undesirable and must be prevented in VLSI circuits.

There are two important mechanisms of reverse breakdown, which are the avalanche and zener mechanisms. Avalanche breakdown occurs in one-sided n^+-p or p^+-n junctions and involves impact ionization [11]. Zener breakdown involves the quantum mechanical tunneling of electrons through an extremely thin depletion layer and, for this reason, can only occur in junctions having very heavy doping on *both* sides of the junction.

In the case of avalanche breakdown, carriers in the depletion region gain significant energy from the high electric field that they ionize silicon lattice atoms on 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 reverse bias. The ionization coefficients that describe the avalanche process are exponential functions of the electric field so that the breakdown voltage decreases strongly with increasing impurity concentration on the lightly doped side of the junction.

Zener breakdown requires that electrons can tunnel through the depletion region under reverse bias conditions. This requires a depletion width on the order of 10 nm or less and can only occur with heavy doping on both sides of the junction.

Figure 3.9 shows the room temperature reverse breakdown voltages for silicon one-sided n^+-p junctions as a function of N_A. These results apply approximately 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 whereby both the avalanche and zener mechanisms contribute.

3.9 Metal-Semiconductor Junctions

Electrical contacts to transistors and other devices necessitate the use of metal-semiconductor junctions. Ideally, the resulting junctions should be able to flow an arbitrary current density with zero voltage drop. Real contacts

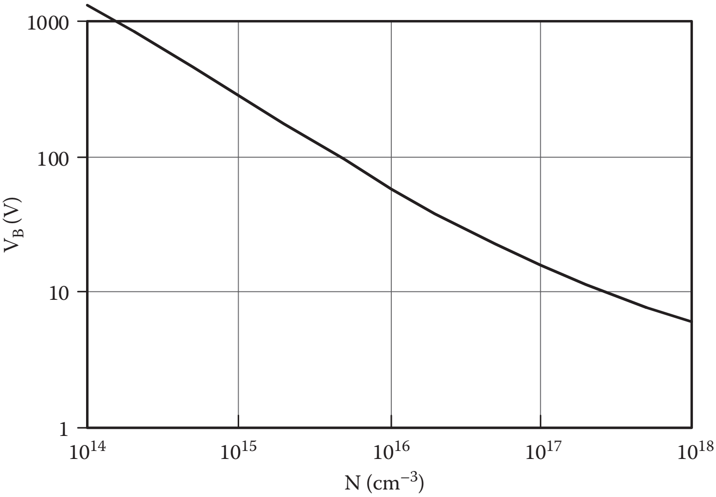


FIGURE 3.9
Reverse breakdown voltage as a function of acceptor concentration for silicon n^+-p junctions at room temperature. The results hold approximately for n-p^+ junctions as long as the donor concentration is used.

that exhibit linear current versus voltage characteristics and low resistance approximate the ideal and are called *ohmic*. Typically, these ohmic contacts are made by producing a junction between a metal and a heavily doped semiconductor. The resulting depletion region is so narrow that electrons may readily tunnel through it, providing ohmic behavior with either polarity of voltage bias. Either a p^+-metal or n^+-metal junction will behave in this manner. Also, because heavily doped p^+-n^+ junctions act as tunnel junctions, it is possible to create multilevel ohmic contact schemes such as p^+-n^+ metal or n^+-p^+ metal. The "metal" used in these contacts may be an elemental metal such as Al, an alloy such as Al-Cu-Si, or a silicide such as $TiSi_2$.

On the other hand, if a metal-semiconductor junction is formed using a lightly doped semiconductor region, the result is a rectifying Schottky junction. This device is useful in clamping, rectifying, or switching applications but undesirable at the contacts to a transistor.

In a metal-n-semiconductor diode, the forward bias current flows from metal to n-semiconductor (the metal acts as the anode), whereas 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). Current flow in a Schottky junction is by thermionic emission rather than minority carrier injection and diffusion. It is therefore a majority carrier device, and the absence of minority carrier storage effects makes it inherently fast. The Schottky junction can still be modeled using the diode equation, but typically the reverse saturation current is orders of magnitude higher than in a p-n junction, resulting in a lower effective *turn-on voltage*.

3.10 SPICE Models

The SPICE model for the p-n junction 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

$$I_D' = IS\left[\exp\left(\frac{qV_D'}{NkT}\right) - 1\right], \tag{3.57}$$

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 junction is entirely attributable to diffusion, then the emission coefficient is unity as predicted in the standard diode equation. At high current densities, the emission coefficient is greater than one as a result of drift-aided diffusion under so-called *high-level injection*. For the purpose of SPICE modeling, this is accounted for by using a modified emission coefficient, using in the range $1 < N < 2$.

The series resistance is accounted for by

$$V_D = V_D' + I_D RS, \tag{3.58}$$

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

$$C_D = TT\frac{IS}{NV_T}\exp\left(\frac{qV_D'}{NkT}\right) + \frac{CJO}{\left(1 - \frac{V_D'}{VJ}\right)^M}, \tag{3.59}$$

where the first term is the diffusion capacitance, attributable to stored excess minority carriers, and the second term is the depletion layer capacitance. For

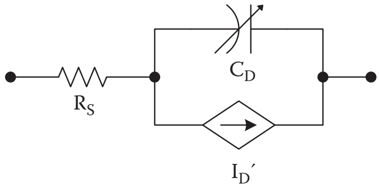


FIGURE 3.10
Circuit model for the p-n junction.

the purpose of SPICE modeling, the parameter TT is called the transit time but is actually equal to the effective forward transit time τ_F . M is the *grading coefficient* and typically varies between ⅓ and ½. For an abrupt junction, M = ½ 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, but there are two important differences. First, the saturation current is orders of magnitude higher than for a p-n junction having 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.

3.11 Practical Perspective

For practical perspective, articles see the dynamic website at http://www.engr.uconn.edu/ece/books/ayers.

3.12 Summary

Digital integrated circuits rely on MOS transistors fabricated in the semiconductor silicon. The silicon is doped selectively with donor impurities (arsenic, phosphorus, and antimony) and the acceptor impurities (boron) to create n-type and p-type regions. Current transport in the resulting devices occurs by drift and diffusion of carriers and is governed by the continuity equation. At low electric field intensities, the carrier drift velocities are

TABLE 3.1

SPICE Parameters for the p-n Junction 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			0.5	0.5

proportional to the electric field. Under the high-field conditions that may exist in short-channel MOS transistors, the carriers approach their saturation velocities.

Semiconductor devices such as MOS transistors involve p-n junctions, in which p-type and n-type material are brought into contact. These p-n junctions are partly responsible for the capacitances, leakage currents, and breakdown voltages in MOS transistors.

3.13 Exercises

E3.1. Determine the room temperature carrier concentrations and resistivity for a silicon wafer doped with boron to a concentration of 10^{15} cm^{-3}.

E3.2. Determine the room temperature carrier concentrations and resistivity for an n-well having a net phosphorus concentration of 10^{16} cm^{-3}.

E3.3. What is the (approximate) maximum electric field intensity for which constant mobility may be assumed for electrons in silicon? In a digital circuit with a supply voltage of 1 V, what is the corresponding gate length for MOS transistors?

E3.4. Repeat E3.3 for the case of holes in silicon.

E3.5. Source and drain regions of an n-MOS transistor are produced by ion implantation of arsenic (~10^{19} cm^{-3}) into a p-type substrate doped to a concentration of 10^{15} cm^{-2}. Find the zero-bias capacitance per square micrometer for the p-n junction.

E3.6. The drain region of a p-MOS transistor is an ion implanted region having a boron concentration of 10^{19} cm^{-3} and a depth of 0.1 μm. This source region is created within an n-well doped to a concentration of 10^{16} cm^{-3} and with a depth of 1.0 μm. Assuming that the source-well p-n junction acts as a short-base diode, estimate the reverse leakage current of this diode based on the diffusion theory. The source area is 2 μm^2.

E3.7. For the source p-n junction described in the previous problem, estimate the reverse leakage current assuming that it is attributable to minority carrier generation rather than diffusion. The space charge lifetime is 10 ns, and the reverse bias is 2 V.

For additional exercise problems, see the dynamic website at http://www.engr.uconn.edu/ece/books/ayers.

References

1. Cohen, E.R., and Taylor, B.N., *The 1986 adjustment to the Fundamental Physical Constants, report of the Committee on Data for Science and Technology of the International Council of Scientific Unions (CODATA) Task Group on Fundamental Constants*, CODATA Bulletin 63, Pergamon, Elmsford, NY, 1986.
2. Cohen, M.L., and Chelikowsky, J.R., *Electronic structure and optical properties of semiconductors*, 2nd ed., Springer-Verlag, Berlin, 1988.
3. Bulucea, C., Recalculation of Irvin's resistivity curves for diffused layers in silicon using updated bulk resistivity data. *Solid-State Electron.*, 36, 489–493, 1993.
4. Jacoboni, C., Canali, C., Ottaviani, G., and Quaranta, A.A., A review of some charge transport properties of silicon. *Solid-State Electron.*, 20, 77–89, 1977.
5. Smith, P., Inoue, M., and Frey, J., Electron velocity in Si and GaAs at very high electric fields. *Appl. Phys. Lett.*, 37, 797–799, 1980.
6. Shockley, W., The theory of p-n junctions in semiconductors and p-n junction transistors. *Bell Sys. Tech. J.*, 28, 435, 1949.
7. Shockley, W., *Electrons and holes in semiconductors*, D. Van Nostrand, Princeton, NJ, 1950.
8. 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, 1228–1243, 1957.
9. Moll, J.L., The evolution of the theory of the current-voltage characteristics of p-n junctions. *Proc. IRE*, 46, 1076–1082, 1958.
10. Ghandhi, S.K., *The theory and practice of microelectronics*, Wiley, New York, 1968.
11. Ghandhi, S.K., *Semiconductor power devices*, Wiley, New York, 1977.

4

The MOS Transistor

4.1 Introduction

The MOSFET is the most important device for digital integrated circuits today. It is a unipolar device, that is, electrical current is carried predominantly by the drift of one type of carrier: electrons in the n-MOS transistor and holes in the p-MOS device. This makes the MOS transistor inherently fast switching, because minority carrier storage effects are unimportant, and the switching speed is limited predominantly by parasitic capacitances. It is a field effect, or voltage-controlled, device, which makes the standby power consumption low. These characteristics, and the ease with which the MOSFET geometry can be scaled down in size, make it very attractive for VLSI circuits.

Figure 4.1 shows the basic structure of an n-MOS transistor. The voltage applied at the gate controls the flow of current between the drain and source. A fourth terminal, the body, may be biased to modify the device characteristics. The polysilicon gate is insulated from the channel region of the device by a thin (~10 nm) layer of silicon dioxide so that, under normal conditions, there is no gate current. When the gate is biased positively with respect to the source, negatively charged electrons are attracted to the interface between the semiconductor and oxide forming an n-type inversion layer. This creates a conducting channel between the drain and the source regions. If the drain is biased positively with respect to the source, electrons in the channel will drift from the source to the drain, thus giving rise to conventional current flow from drain to source. The operation of a p-MOS transistor is qualitatively similar, but the source, drain, and inversion layer are all p-type. Application of negative voltages on the gate and drain with respect to the source will cause holes to drift from the source to drain, with an associated conventional current in the same direction.

Figure 4.1 displays an enhancement-type n-MOS transistor; for this device, there is no conducting channel between the drain and source unless a positive voltage is applied between the gate and source so it is a normally-off

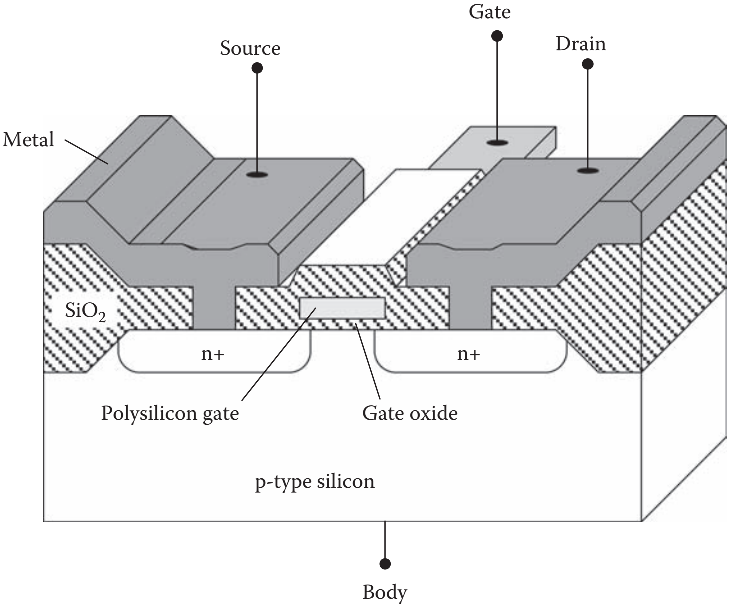


FIGURE 4.1
MOSFET.

transistor. (The name *enhancement type* reflects the fact that a gate bias is required to enhance a conducting channel.) Some MOS transistors are designed to conduct with zero gate-source bias, and these are referred to as *depletion type* devices. However, enhancement type devices are preferred in digital circuits for low standby power.

Various MOSFET symbols have been used in the literature. Some of the most common ones are shown in Figures 4.2 and 4.3, for enhancement-type n-MOS and p-MOS transistors, respectively. Throughout this book, the symbols depicted in Figures 4.2a and 4.3a will generally be used, except in those cases in which it is necessary to show the connection to the transistor body.

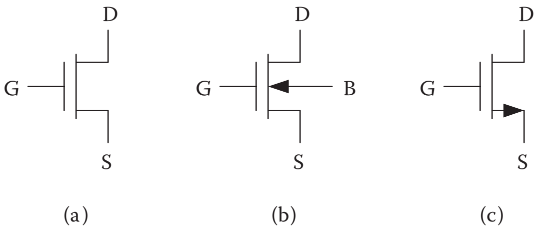


FIGURE 4.2
Symbols for enhancement-type n-MOS transistors.

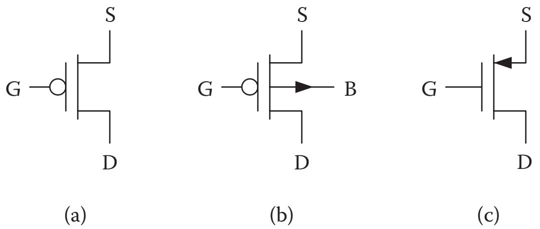


FIGURE 4.3
Symbols for enhancement-type p-MOS transistors.

4.2 The 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 as well as the semiconductor, by the bias on a metal gate. Modulation of the gate bias voltage allows control of the channel conductivity and therefore the drain current. This behavior may be understood by consideration of an MOS capacitor [1–3] as shown in Figure 4.4. In this figure, the p-type substrate has been grounded and a sufficiently positive gate bias has been applied so that an inversion layer of electrons is established under the oxide. Under this condition, a depletion region exists beneath the inversion layer. Here the material is depleted of mobile holes, although it has not been inverted to n-type conductivity. For a more quantitative description of MOSFET operation, it is necessary to determine the threshold voltage V_T that will give rise to an inversion layer in the semiconductor.

The starting point for this analysis is to consider the case of an ideal* MOS capacitor with zero bias as shown in Figure 4.5. For specificity, a p-type substrate has been assumed, but the case of an n-type substrate may be analyzed in similar manner. In this thermal equilibrium (zero bias) condition, 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 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 $q\phi_s$ depends on the semiconductor electron affinity and also the doping in the semiconductor. The separation between the intrinsic Fermi level E_i and the Fermi level E_f in the semiconductor is $q\varphi_F$. For the ideal case in which the work functions of the metal and semiconductor are equal and there is no net electrical charge in the oxide, zero applied bias results in the flat band

* In the ideal MOS capacitor, (1) there is zero work function difference between the metal and the semiconductor, and (2) there are no electrical charges in the oxide.

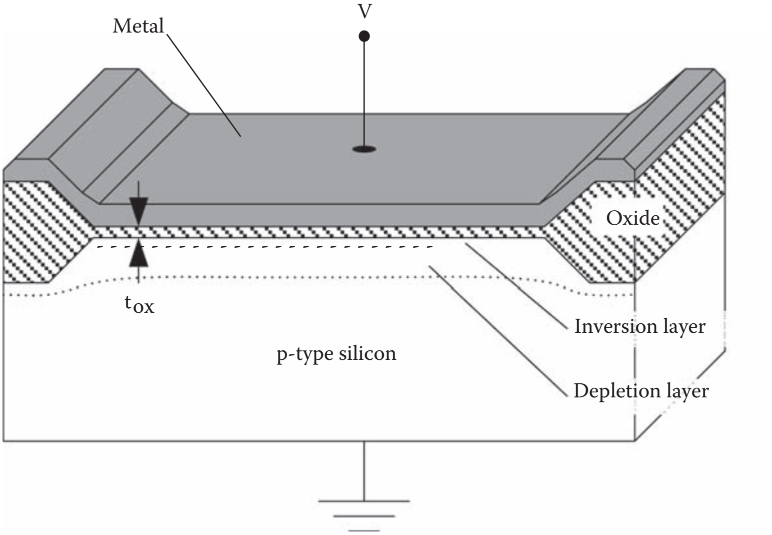


FIGURE 4.4
MOS capacitor.

condition depicted in Figure 4.5. With flat bands in the semiconductor, the hole concentration is the same throughout the depth of the p-type silicon.

Application of a positive bias on the metal gate with respect to the semiconductor will give rise to band bending in the semiconductor as shown in Figure 4.6. The extent of the band bending is $q\psi_s$, and 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 with the bulk of the p-type silicon. The hole concentration is consequently reduced near the interface, giving rise to a *depletion* condition.

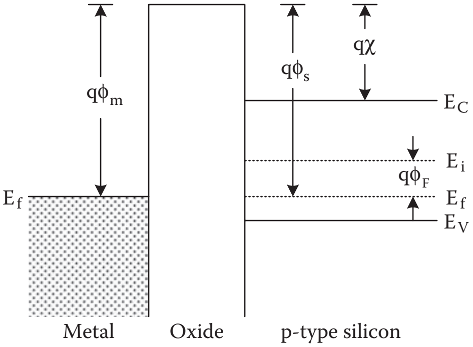


FIGURE 4.5
Ideal MOS capacitor under the flat-band condition.

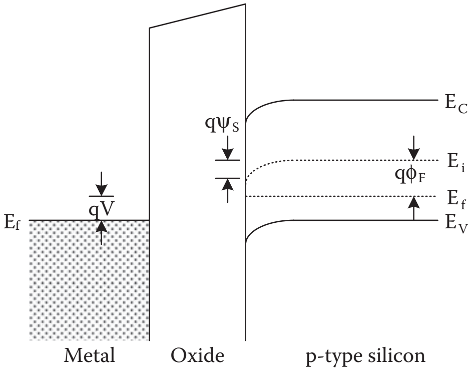


FIGURE 4.6
Ideal MOS capacitor under the depletion condition.

The application of a sufficiently positive bias on the gate will result in inversion. In this case, the band bending is enough that the semiconductor becomes n-type near the interface. *Inversion* refers to the condition in which the semiconductor is simply converted to n-type near the interface, but *strong inversion* (depicted in Figure 4.7) describes the case in which the electron concentration is equal to or greater than the hole concentration deep below the gate (the hole concentration in the bulk of the semiconductor). It is customary to assume that the condition of strong inversion corresponds to the threshold voltage for conduction in the MOSFET, and this occurs when the band bending in the semiconductor is given by

$$q\psi_S = 2q\phi_F \,. \tag{4.1}$$

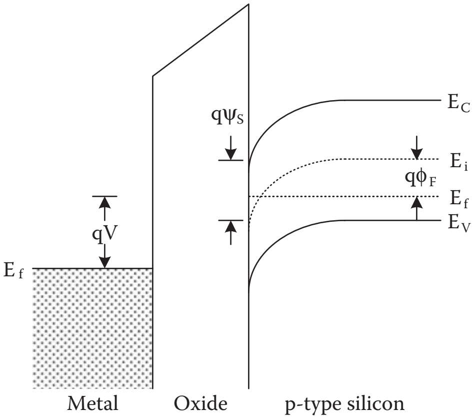


FIGURE 4.7
Ideal MOS capacitor under the condition of strong inversion.

In an ideal MOS capacitor, the voltage bias that results in strong inversion is

$$V_{inv} = 2\phi_F - \frac{Q_B}{C_{ox}}, \tag{4.2}$$

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 per unit area,

$$C_{ox} = \frac{\varepsilon_{ox}}{t_{ox}}, \tag{4.3}$$

and Q_B is the charge per unit area in the depletion layer of the semiconductor under strong inversion, given by

$$Q_B = -\sqrt{2\varepsilon_{Si} q N_a \left|2\phi_F\right|}\ . \tag{4.4}$$

Equation 4.4 applies to the ideal MOS capacitor, but, in real devices, we must consider the work function difference and the oxide charge.

4.3 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 the semiconductor, and the oxide charge, we find the threshold voltage to be

$$V_{TO} = \phi_{MS} - 2\phi_F - \frac{Q_B}{C_{ox}} - \frac{Q_{ox}}{C_{ox}} - \frac{Q_{II}}{C_{ox}}, \tag{4.5}$$

where φ_{MS} is the function difference between metal and semiconductor, $2\varphi_F$ is the voltage across the semiconductor necessary to create a conducting channel (inversion layer), Q_B is the charge per unit area in the semiconductor under inversion, Q_{ox} is the charge per unit area in oxide, Q_{II} is the charge per unit area of ion implanted impurities in the semiconductor, and C_{ox} is the oxide capacitance per unit area.

For an n-channel MOSFET with an acceptor-doped channel region,

$$\phi_F = \frac{kT}{q}\ln\left(\frac{n_i}{N_a}\right). \tag{4.6}$$

If the gate is assumed to be heavily doped polysilicon (*degenerate* polysilicon), then the work function difference is

$$\phi_{MS} = \frac{kT}{q}\ln\left(\frac{n_i}{N_a}\right) - \frac{E_g}{2q}, \tag{4.7}$$

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

$$\frac{Q_B}{C_{ox}} = \frac{-\sqrt{2qN_a\varepsilon_{Si}\left|2\phi_F\right|}}{\varepsilon_{ox} / t_{ox}}. \tag{4.8}$$

If there is a non-zero-bias V_{BS} applied between the body and the source of the MOSFET, this further modifies the threshold voltage to

$$V_T = V_{TO} + \gamma\left(\sqrt{\left|V_{BS} + 2\phi_F\right|} - \sqrt{\left|2\phi_F\right|}\right), \tag{4.9}$$

where γ is the body effect coefficient, given by

$$\gamma = \frac{\sqrt{2q\varepsilon_{Si}N_a}}{C_{ox}}, \tag{4.10}$$

where q is the electronic charge (1.602×10^{-19} C), ε_{Si} is the permittivity of silicon, N_a is the acceptor doping in the p-type silicon (n-channel MOSFET), and C_{ox} is the oxide capacitance per unit area.

An important 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. The use of arsenic or phosphorus makes the threshold voltage more negative, whereas the implantation of boron makes it more positive. The actual threshold voltage can be of either sign, but for enhancement-type transistors normally used in VLSI circuits, the n-MOS transistors have positive threshold voltages, whereas the p-MOS transistors have negative threshold voltages.

Even after fabrication, the threshold of a MOSFET may be adjusted *in the circuit* using the body bias effect. 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 used in modern low-power, high-speed CMOS circuits.

Example 4.1 Threshold Voltage for n-MOSFET

Calculate the zero-bias threshold voltage for an n-channel MOSFET with $t_{ox} = 10nm$ 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 there are $10^{11}cm^{-2}$ positive charges in the oxide. A boron dose of $1.3 \times 10^{12}cm^{-2}$ is implanted to adjust the threshold voltage.

Solution: The zero-bias threshold is

$$V_{TO} = \phi_{MS} - 2\phi_F - \frac{Q_B}{C_{ox}} - \frac{Q_{ox}}{C_{ox}} - \frac{Q_{II}}{C_{ox}}.$$

The work function difference is

$$\phi_{MS} = \frac{kT}{q}\ln\left(\frac{n_i}{N_a}\right) - \frac{E_g}{2q} = -0.35V - \frac{1.12V}{2} = -0.90V.$$

For this level of doping,

$$\phi_F = \frac{kT}{q}\ln\left(\frac{n_i}{N_a}\right) = (0.026V)\ln\left(\frac{1.45\times10^{10}cm^{-3}}{10^{16}cm^{-3}}\right) = -0.35V.$$

The oxide capacitance per unit area is

$$C_{ox} = \frac{\varepsilon_{ox}}{t_{ox}} = \frac{(3.9)(8.85\times10^{-14}F/cm)}{10\times10^{-7}cm} = 3.45\times10^{-7}F/cm^2.$$

The depletion charge per unit area under strong inversion is

$$\begin{aligned} Q_B &= -\sqrt{2qN_a\varepsilon_{Si}\left|2\phi_F\right|} \\ &= -\sqrt{2(1.602\times10^{-19}C)(10^{16}cm^{-3})(11.9)(8.85\times10^{-14}F/cm)\left|2(-0.35V)\right|} \\ &= -4.86\times10^{-8}C/cm^2. \end{aligned}$$

The oxide charge per unit area is

$$Q_{ox} = (1.602\times10^{-19}C)(10^{11}cm^{-2}) = 1.60\times10^{-8}C/cm^2.$$

The boron ion implantation charge per unit area is

$$Q_{II} = -(1.602\times10^{-19}C)(1.3\times10^{12}cm^{-2}) = -2.10\times10^{-7}C/cm^2.$$

Therefore, the zero-bias threshold voltage is

$$\begin{aligned} V_{TO} &= \phi_{MS} - 2\phi_F - \frac{Q_B}{C_{ox}} - \frac{Q_{ox}}{C_{ox}} - \frac{Q_{II}}{C_{ox}} \\ &= -0.90V - (-0.70V) - \frac{-4.86\times10^{-8}C/cm^2}{3.45\times10^{-7}F/cm^2} - \frac{1.60\times10^{-8}C/cm^2}{3.45\times10^{-7}F/cm^2} - \frac{-2.10\times10^{-7}C/cm^2}{3.45\times10^{-7}F/cm^2} \\ &= -0.90V + 0.70V + 0.14V - 0.05V + 0.61V \\ &= 0.50V. \end{aligned}$$

The ion implantation adjustment was necessary to obtain an enhancement-type transistor with a positive threshold voltage.

Example 4.2 Threshold Voltage for p-MOSFET

A p-MOS transistor is fabricated with $t_{ox} = 10nm$ and $N_d = 10^{16}cm^{-3}$ in the n-well. The gate is heavily doped p-polysilicon, and there are $5 \times 10^{10}cm^{-2}$ positive charges in the oxide. Determine the necessary ion implantation (type and dose) to adjust the zero-bias threshold voltage to −0.5V.

Solution: After ion implantation, the zero-bias threshold is

$$V_{TO} = \phi_{MS} - 2\phi_F - \frac{Q_B}{C_{ox}} - \frac{Q_{ox}}{C_{ox}} - \frac{Q_{II}}{C_{ox}}.$$

The work function difference for the p-MOS transistor is

$$\phi_{MS} = \frac{kT}{q}\ln\left(\frac{N_d}{n_i}\right) + \frac{E_g}{2q} = 0.35V + \frac{1.12V}{2} = 0.90V \cdot$$

For the n-well,

$$\phi_F = \frac{kT}{q}\ln\left(\frac{N_d}{n_i}\right) = (0.026V)\ln\left(\frac{10^{16}cm^{-3}}{1.45 \times 10^{10}cm^{-3}}\right) = 0.35V \cdot$$

The oxide capacitance per unit area is

$$C_{ox} = \frac{\varepsilon_{ox}}{t_{ox}} = \frac{(3.9)(8.85 \times 10^{-14}F / cm)}{10 \times 10^{-7}cm} = 3.45 \times 10^{-7}F / cm^2 \cdot$$

The depletion charge per unit area under strong inversion is

$$\begin{aligned} Q_B &= \sqrt{2qN_d\varepsilon_{Si}\left|2\phi_F\right|} \\ &= \sqrt{2(1.602 \times 10^{-19}C)(10^{16}cm^{-3})(11.9)(8.85 \times 10^{-14}F / cm)\left|2(0.35V)\right|} \\ &= 4.86 \times 10^{-8}C / cm^2. \end{aligned}$$

The oxide charge per unit area is

$$Q_{ox} = (1.602 \times 10^{-19}C)(5 \times 10^{10}cm^{-2}) = 8.01 \times 10^{-9}C / cm^2.$$

Before ion implantation, the zero-bias threshold voltage is

$$\begin{aligned} V_{TO} &= \phi_{MS} - 2\phi_F - \frac{Q_B}{C_{ox}} - \frac{Q_{ox}}{C_{ox}} \\ &= 0.90V - (0.70V) - \frac{4.86 \times 10^{-8}C / cm^2}{3.45 \times 10^{-7}F / cm^2} - \frac{8.01 \times 10^{-9}C / cm^2}{3.45 \times 10^{-7}F / cm^2} \\ &= 0.90V - 0.70V - 0.14V - 0.02V \\ &= 0.04V. \end{aligned}$$

To adjust the threshold to −0.5V, the necessary ion implantation charge per unit area is

$$Q_{II} = -\Delta V_{TO} C_{ox} = -\left(-0.50V - 0.04V\right)\left(3.45\times10^{-7} F / cm^2\right) = 1.86\times10^{-7} C / cm^2.$$

The positive ion implantation charge indicates the use of a donor such as phosphorus, and the necessary dose may be found from

$$N_{II} = \frac{|Q_{II}|}{q} = \frac{1.86\times10^{-7} C / cm^2}{1.602\times10^{-19} C} = 1.16\times10^{12} cm^{-2} .$$

Example 4.3 Body Bias Effect in n-MOSFET

Consider an n-MOS transistor with $t_{ox} = 10nm$ and $N_a = 10^{16} cm^{-3}$. Calculate the change in the threshold voltage attributable to the body bias effect if $V_{BS} = -1.5V$.

Solution: The body effect coefficient is

$$\begin{aligned}\gamma &= \frac{\sqrt{2q\varepsilon_{Si} N_a}}{C_{ox}} \\ &= \frac{\sqrt{2\left(1.602\times10^{-19} C\right)(11.9)\left(8.85\times10^{-14} F / cm\right)\left(10^{16} cm^{-3}\right)}}{(3.9)\left(8.85\times10^{-14} F / cm\right) / 10\times10^{-7} cm} \\ &= 0.168V^{1/2}.\end{aligned}$$

For the p-substrate,

$$\phi_F = \frac{kT}{q}\ln\left(\frac{n_i}{N_a}\right) = (0.026V)\ln\left(\frac{1.45\times10^{10} cm^{-3}}{10^{16} cm^{-3}}\right) = -0.35V .$$

Therefore, the change in the threshold voltage will be

$$\begin{aligned}\Delta V_T &= V_T - V_{TO} \\ &= \gamma\left(\sqrt{|V_{BS} + 2\phi_F|} - \sqrt{|2\phi_F|}\right) \\ &= \left(0.168V^{1/2}\right)\left(\sqrt{2.20V} - \sqrt{0.70V}\right) \\ &= 0.109V.\end{aligned}$$

Here the negative body bias renders the threshold voltage more positive. (In the case of a p-MOS transistor, a positive body bias makes the threshold voltage more negative.)

4.4 MOSFET Current-Voltage Characteristics

Consider an n-channel enhancement-type MOSFET biased as in Figure 4.8. Here V_{GS} is the gate-to-source bias, V_{DS} is the drain-to-source bias, and I_D is the drain current. We will assume that the p-type substrate has been connected to the source so that $V_{BS} = 0$. (In general, the p-type substrate is tied to ground, whereas the n-well is tied to $+V_{DD}$. This means that, in NAND and NOR gates, the body-source bias will not always be zero for all devices.)

There are three modes of operation for the MOSFET: *cutoff*, *linear*, and *saturation*. *Cutoff* occurs if the gate-to-source bias voltage is less than the threshold voltage and is also referred to as *subthreshold* operation. As a first-order approximation, cutoff results in zero drain current. (However, real devices exhibit subthreshold currents that are important in dynamic circuits and low-power applications.) When 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. In the case of a small drain-to-source bias ($V_{DS} \le V_{GS} - V_T$), the MOSFET acts like a voltage-controlled resistance,

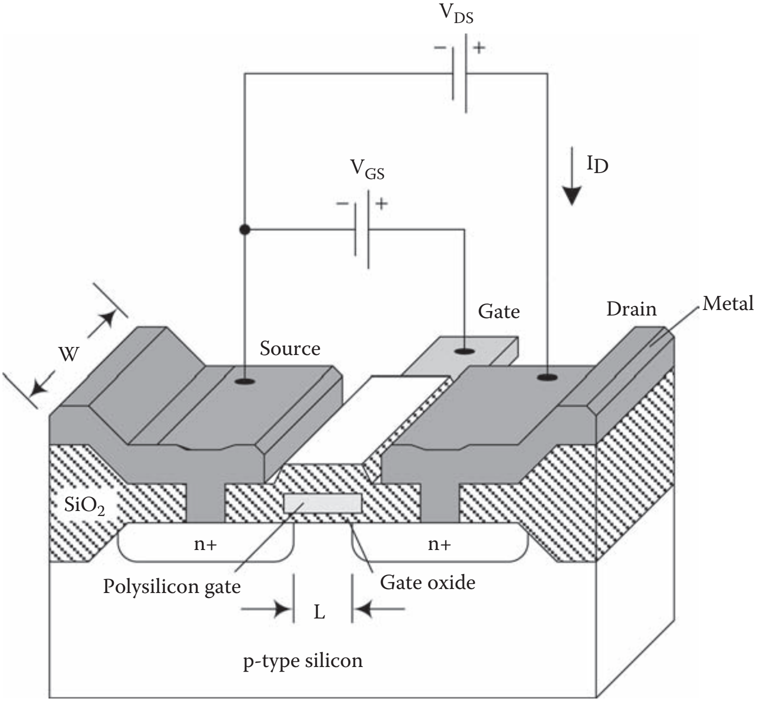


FIGURE 4.8
Enhancement-type n-MOS transistor with bias.

and this mode of operation is called the *linear* or *ohmic* mode of operation. With a larger drain-to-source bias ($V_{DS} \geq V_{GS} - V_T$), the conducting channel pinches off at the drain end, causing the drain current to saturate; this mode of operation is called *saturation*.

The characteristic curves for a MOSFET are shown qualitatively in Figure 4.9. These illustrate the drain current I_D versus the drain-to-source voltage V_{DS} with the gate-to-source voltage V_{GS} as a parameter. The result is 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, left of the parabola, the drain current increases *approximately linearly* with the drain-to-source voltage. Saturation is characterized by *approximately constant* drain current, with 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.

4.4.1 Linear Operation

Linear operation occurs if the gate-to-source bias is more positive than the threshold ($V_{GS} \geq V_T$), 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. Table 4.1 summarizes the voltage boundaries for the linear region of operation.

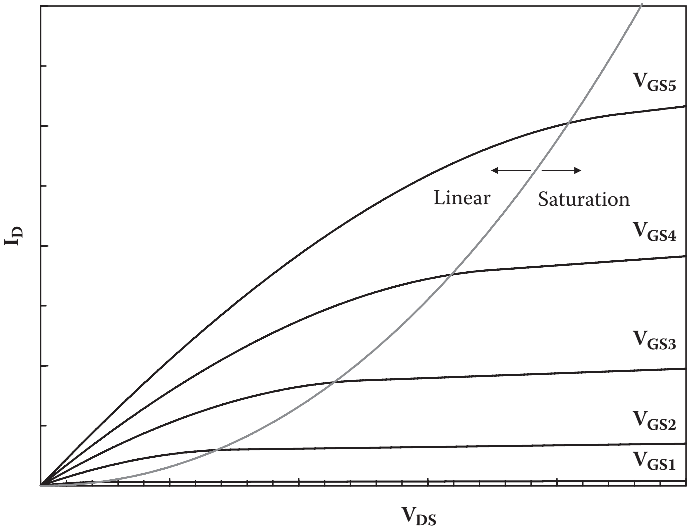


FIGURE 4.9
Characteristic curves for a MOSFET.

TABLE 4.1

Voltage Boundaries for Linear Operation of MOSFETs

	Linear operation	
n-MOS	$V_{GS} \geq V_T$	$V_{DS} \leq (V_{GS} - V_T)$
p-MOS	$V_{GS} \leq V_T$	$V_{DS} \geq (V_{GS} - V_T)$

The drain current in an n-MOS transistor with linear operation may be determined with the aid of Figure 4.10. For the assumed coordinate system, the origin is at the source end of the channel, y is the distance along the channel parallel to the oxide interface, and x is the perpendicular distance from the oxide interface.

Our starting point is the assumption that carriers move by drift only and that the drift is only in the y direction parallel to the interface (the *gradual channel approximation*) [4]. Based on this assumption, the drain current at a point y along the channel is

$$I_D(y) = qW \int_{x=0}^{\infty} \mu_n n(x,y) \frac{dV}{dy} dx \,. \tag{4.11}$$

Integration over x (the depth of the inversion layer of conducting electrons) yields

$$I_D(y) = -\mu_n W \frac{dV}{dy} Q_i(y) \,, \tag{4.12}$$

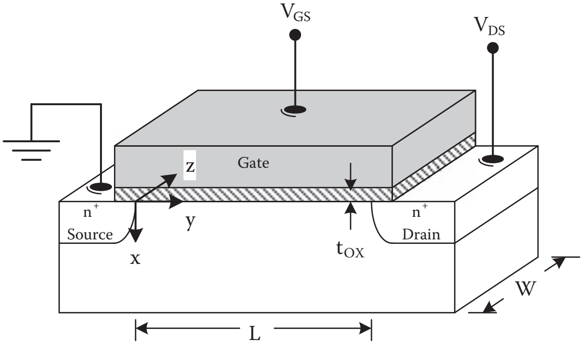


FIGURE 4.10
MOSFET structure for determination of the drain current.

where $Q_i(y)$ is the integrated electron charge in the inversion layer per unit area at a distance along the channel y. If we assume that the inversion layer charge is located in a sheet of zero thickness (the *charge sheet approximation*) [5] and that the bulk depletion charge is approximately independent of the gate-to-source bias, then

$$Q_i\left(y\right) \approx -\frac{\varepsilon_{ox}}{t_{ox}}\left(V_{GS} - V\left(y\right) - V_T\right). \tag{4.13}$$

Therefore,

$$I_D dy = \frac{\mu_n \varepsilon_{ox}}{t_{ox}} W\left(V(y) - V_T\right) dV \ . \tag{4.14}$$

Integrating over the length of the channel, we obtain

$$\int_0^L I_D dy = \frac{\mu_n \varepsilon_{ox}}{t_{ox}} W \int_0^{V_{DS}} \left(V_{GS} - V(y) - V_T\right) dV \ . \tag{4.15}$$

This provides an equation for the drain current in the linear region of operation,

$$I_D = \frac{\mu_n \varepsilon_{ox} W}{t_{ox} L}\left[\left(V_{GS} - V_T\right)V_{DS} - \frac{V_{DS}^2}{2}\right] . \tag{4.16}$$

The result above 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, carrier diffusion can be significant near the threshold, and, in short-channel transistors, the high electric field intensity can result in carrier velocity saturation and must be considered. Second, the variation of the depletion charge with the gate-to-source bias has been neglected compared with the inversion charge, but this is only valid for lightly doped channel regions.

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

$$V_{GS} - V_{DS} = V_T \ , \tag{4.17}$$

and 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* K :

$$I_D = K\left[\left(V_{GS} - V_T\right)V_{DS} - \frac{V_{DS}^2}{2}\right], \qquad V_{GS} \geq V_T \text{ and } \left(V_{GS} - V_T\right) \geq V_{DS}\, , \tag{4.18}$$

in which the device transconductance parameter is

$$K = \frac{W}{L}\frac{\mu_n \varepsilon_{ox}}{t_{ox}}, \tag{4.19}$$

where W is the width of the device, L is the length of the device, μ_n is the mobility of electrons (n-channel device), ε_{ox} is the permittivity of oxide, and t_{ox} is oxide thickness.

Sometimes the device transconductance parameter is calculated as

$$K = \frac{W}{L}k^{'}, \tag{4.20}$$

where $k^{'}$ is the *process transconductance parameter* given by

$$k^{'} = \frac{\mu_n \varepsilon_{ox}}{t_{ox}} = \mu_n C_{ox}. \tag{4.21}$$

Another useful relationship for linear operation allows the calculation of the drain-to-source voltage, if the gate-to-source voltage is known and the drain current is known. If the channel length modulation is neglected, then use of the quadratic formula yields

$$V_{DS} = (V_{GS} - V_T) - \sqrt{(V_{GS} - V_T)^2 - \frac{2I_D}{K}}, \quad V_{GS} \geq V_T \text{ and } (V_{GS} - V_T) \geq V_{DS}. \tag{4.22}$$

The relationships for p-channel MOSFETs may be obtained in a 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 4.4 Process Transconductance Parameters

Estimate the process transconductance parameters for n-channel and p-channel MOSFETs with $t_{ox} = 10\text{nm}$.

Solution: For n-channel MOSFETs, assuming $\mu_n = 580cm^2V^{-1}s^{-1}$, the process transconductance parameter is

$$k_N^{'} = \frac{\mu_n \varepsilon_{ox}}{t_{ox}} = \frac{(580cm^2V^{-1}s^{-1})(3.9)(8.85\times10^{-14}F/cm)}{10\times10^{-7}cm} = 200\mu A/V^2.$$

For p-channel MOSFETs, assuming $\mu_p = 230cm^2V^{-1}s^{-1}$, the process transconductance parameter is

$$k_P^{'} = \frac{\mu_p \varepsilon_{ox}}{t_{ox}} = \frac{(230cm^2V^{-1}s^{-1})(3.9)(8.85\times10^{-14}F/cm)}{10\times10^{-7}cm} = 79\mu A/V^2.$$

The process transconductance parameter is typically 2.5 times larger for the n-channel MOSFETs compared with p-channel MOSFETs.

Example 4.5 Device Transconductance Parameters

Estimate the device transconductance parameters for n-channel and p-channel MOSFETs with $t_{ox} = 10\text{nm}$, $W = 2\mu m$, and $L = 0.25\mu m$

Solution: For the n-channel MOSFET, the device transconductance parameter is

$$K_N = \frac{W}{L} k_N^{'} = \left(\frac{2\mu m}{0.25\mu m} \right) 200\mu A / V^2 = 1.60 mA / V^2 .$$

For the p-channel MOSFET, the device transconductance parameter is

$$K_P = \frac{W}{L} k_P^{'} = \left(\frac{2\mu m}{0.25\mu m} \right) 80\mu A / V^2 = 0.64 mA / V^2 .$$

If n- and p-channel MOSFETs are fabricated with the same aspect ratio, the n-channel MOSFETs have a device transconductance parameter that is 2.5 times larger.

4.4.2 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

$$V_{DS} = V_{GS} - V_T . \tag{4.23}$$

Substituting this result in the equation for linear operation, we obtain the equation for saturated operation of the MOSFET:

$$I_D = \frac{K}{2} \left(V_{GS} - V_T \right)^2 ; \qquad V_{GS} \geq V_T \text{ and } \left(V_{GS} - V_T \right) \leq V_{DS} . \tag{4.24}$$

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.

Tables 4.2 and 4.3 summarize the long-channel drain current equations for n-MOS and p-MOS transistors.

4.4.3 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.

TABLE 4.2

Drain Current Equations for a Long-Channel n-MOS Transistor

Mode	Drain current equation	Voltage conditions
Cutoff	$I_D \approx 0$	$(V_{GS} - V_{TN}) \le 0$
Linear	$I_D = \mu_n C_{ox} \frac{W}{L}\left[(V_{GS} - V_{TN})V_{DS} - \frac{V_{DS}^2}{2}\right]$	$V_{DS} \le (V_{GS} - V_{TN})$
Saturation	$I_D = \mu_n C_{ox} \frac{W}{L}\frac{(V_{GS} - V_{TN})^2}{2}$	$0 \le (V_{GS} - V_{TN}) \le V_{DS}$

Typically, $\mu_n = 580 cm^2 / Vs$. The drain current is assumed to be entering the drain terminal. Normally, V_{DS}, V_{GS}, and V_{TN} are all positive.

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 of importance in modern low-voltage, low-power CMOS and memory circuits.

Whereas 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 [3]. 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 (non-inverted) p-type channel region and diffuse to the drain, giving rise to a conventional current flow from the drain to the source.

If it is assumed that diffusion alone is responsible for the subthreshold current, then

$$I_D = -qAD_n \frac{dn}{dy} = qAD_n \frac{n(0) - n(L)}{L}, \tag{4.25}$$

TABLE 4.3

Drain Current Equations for a Long-Channel p-MOS Transistor

Mode	Drain current equation	Voltage conditions
Cutoff	$I_D \approx 0$	$(V_{GS} - V_{TP}) \ge 0$
Linear	$I_D = \mu_p C_{ox} \frac{W}{L}\left[(V_{GS} - V_{TP})V_{DS} - \frac{V_{DS}^2}{2}\right]$	$V_{DS} \ge (V_{GS} - V_{TP})$
Saturation	$I_D = \mu_p C_{ox} \frac{W}{L}\frac{(V_{GS} - V_{TP})^2}{2}$	$0 \ge (V_{GS} - V_{TP}) \ge V_{DS}$

Typically, $\mu_p = 230 cm^2 / Vs$. The drain current is assumed to be leaving the drain terminal. Normally, V_{DS}, V_{GS}, and V_{TP} are all negative.

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

$$n(0) \approx \bar{n}_{po} \exp\left(\frac{q\psi_s}{kT}\right), \tag{4.26}$$

and

$$n(L) \approx \bar{n}_{po} \exp\left(\frac{q(\psi_s - V_{DS})}{kT}\right), \tag{4.27}$$

where ψ_s is the band bending in the semiconductor. The effective thickness of the inversion layer is kT / qE_s , where Es is the surface electric field intensity. Thus,

$$A \approx \frac{kTW}{qE_S} = \frac{kTW}{q}\sqrt{\frac{\varepsilon_{Si}}{2qN_a\psi_s}}, \tag{4.28}$$

where the band bending in the semiconductor is given by

$$\psi_s \approx 2\psi_B + \frac{(V_{GS} - V_T)}{m}. \tag{4.29}$$

The unitless parameter m is given by

$$m = 1 + \frac{C_{dm}}{C_{ox}}, \tag{4.30}$$

where C_{dm} is the maximum capacitance of the depletion layer under the oxide per unit area, and C_{ox} is the oxide capacitance per unit area.

Combining the previous equations and making use of the Einstein relationship, we obtain the subthreshold current in the n-MOSFET:

$$\begin{aligned} I_D &= \mu_n \frac{W}{L}\sqrt{\frac{\varepsilon_{Si} q N_a}{4\psi_B}}\left(\frac{kT}{q}\right)^2 \exp\left(\frac{q(V_{GS} - V_T)}{mkT}\right)\left[1 - \exp\left(-\frac{qV_{DS}}{kT}\right)\right] \\ &= \frac{\mu_n \varepsilon_{ox} W(m-1)}{t_{ox}L}\left(\frac{kT}{q}\right)^2 \exp\left(\frac{q(V_{GS} - V_T)}{mkT}\right)\left[1 - \exp\left(-\frac{qV_{DS}}{kT}\right)\right] \end{aligned}. \tag{4.31}$$

In typical MOSFETs, 1 < m < 2.*

* 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 compared with the inversion layer charge. However, this approximation is not appropriate for the subthreshold analysis.

If the drain-to-source bias is several times kT/q ~ 26 mV at room temperature, then the subthreshold current is independent of the drain-to-source bias:

$$I_D \approx K(m-1)\left(\frac{kT}{q}\right)^2 \exp\left(\frac{q(V_{GS}-V_T)}{mkT}\right). \tag{4.32}$$

An important figure for subthreshold operation is the *subthreshold swing*, defined as

$$S \equiv \left(\frac{d(\log_{10} I_D)}{dV_{GS}}\right)^{-1}. \tag{4.33}$$

From Equation 4.32, the subthreshold swing is

$$S = 2.3\frac{mkT}{q}. \tag{4.34}$$

Typically, room temperature operation of MOSFETs is characterized by a subthreshold swing of 100 mV, meaning 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 the 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 8, this is a significant issue in the design of low-power CMOS circuits.

Example 4.6 Subthreshold Current in n-MOSFET

Calculate and plot the subthreshold drain current for an n-channel MOSFET with $K = 2mA/V^2$, $V_T = 0.3V$, and $m = 1.6$.

Solution: Assuming $V_{DS} > 3kT/q$, the subthreshold current is given by

$$\begin{aligned} I_D &\approx K(m-1)\left(\frac{kT}{q}\right)^2 \exp\left(\frac{q(V_{GS}-V_T)}{mkT}\right) \\ &= (2.0mA/V^2)(0.6)(26mV)^2 \exp\left(\frac{V_{GS}-V_T}{(1.6)(26mV)}\right) \\ &= (0.81\mu A)\exp\left(\frac{V_{GS}-V_T}{41mV}\right). \end{aligned}$$

The results are plotted in Figure 4.11 along with the above-threshold saturation current for $V_{GS} > 0.3$ V. In the subthreshold regime, the characteristic is a straight line on a semilog plot, showing that the subthreshold current increases exponentially with the gate-to-source bias.

4.4.4 Transit Time

It takes a finite time for the majority carriers to traverse the channel in a conducting MOSFET. This delay is called the transit time t_t. For the purpose of estimating propagation delays in MOS circuits, we usually assume that the transit time is 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

$$E \approx \frac{V_{DS}}{L}. \tag{4.35}$$

Therefore, the carriers move at a velocity of approximately

$$v \approx \frac{\mu_n V_{DS}}{L}, \tag{4.36}$$

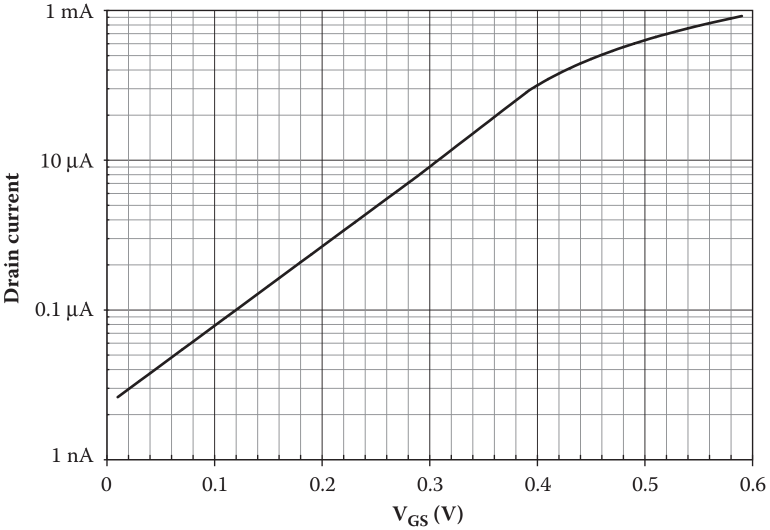


FIGURE 4.11
Drain current as a function of the gate-to-source voltage, for subthreshold to above-threshold operation.

so the transit time is

$$t_t = \frac{L}{v} = \frac{L^2}{\mu_n V_{DS}} \quad . \tag{4.37}$$

Therefore, the transit time increases with the square of the channel length.

Example 4.7 Transit Time in n-MOSFET

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 cm^2/Vs for electrons, the transit time is

$$t_t = \frac{L^2}{\mu_n V_{DS}} = \frac{\left(5\times 10^{-4}\,cm\right)^2}{\left(580 cm^2 V^{-1} s^{-1}\right)\left(5V\right)} = 86ps \; .$$

The quasi-static approximation is applicable if the circuit propagation delay is greater than the transit time.

4.5 Short-Channel MOSFETs

Aggressive scaling of MOSFETs, and in particular the channel lengths in MOSFETs, has resulted in devices that behave differently to the long-channel devices described in Section 4.4. First, the threshold voltage becomes a function of the channel length and width. (These phenomena are called the SCE and the NCE.) Second, the subthreshold characteristics may be degraded, so that the subthreshold current varies significantly with the drain-to-source voltage, as a consequence of DIBL. Third, the electric field intensity in the channel may be sufficiently large so that the carriers reach their saturated velocity. Fourth, the effective channel length becomes a function of the drain-to-source bias because of *channel length modulation*. All of these effects are of practical importance in the design of high-performance CMOS circuits today.

4.5.1 The Short-Channel Effect

As a consequence of the 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 [6] as illustrated in Figure 4.12.

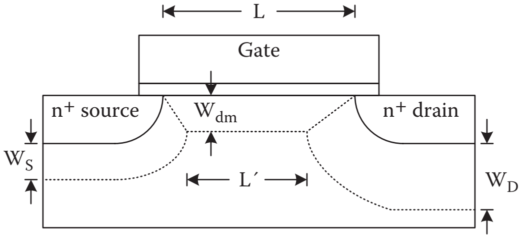


FIGURE 4.12
Charge sharing and the SCE in a short-channel MOSFET.

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 with long-channel value, by

$$\Delta V_T \approx \frac{qN_a t_{ox}}{\varepsilon_{ox}}\left(\frac{L'-L}{2L}\right)W_{dm} \approx \frac{-qN_a t_{ox}}{\varepsilon_{ox}}\left(\frac{W_S+W_D}{2L}\right)W_{dm}, \qquad (4.38)$$

where ε_{ox} is the permittivity of oxide, t_{ox} is oxide thickness, W is the width of MOSFET channel, L is the length of the MOSFET channel, W_{dm} is the depletion width in semiconductor under inversion, W_S is the source junction depletion width, and W_D is the drain junction depletion width.

In practice, careful design can overcome the SCE 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. This requires that the implantation adjustment be designed so that the narrowest devices on the wafer have acceptable threshold voltages.

4.5.2 Narrow-Channel Effect

Another small geometry effect is the NCE, by which *the threshold voltage becomes a function of the device width*. This occurs because the depletion region under the channel extends beyond the width of the gate on either side of the device. Because the gate-to-source bias must support this extra charge, the threshold voltage is increased in a narrow-channel device. This offsets the threshold voltage reduction attributable to the SCE to some extent. However, the NCE can be problematic because design considerations may require significant variations of device widths on a wafer.

4.5.3 Drain-Induced Barrier Lowering

The subthreshold characteristics of a short-channel MOS transistor may degrade as a result of the DIBL [7]. In the subthreshold region of operation, charge carriers moving from the source to the drain experience a potential barrier at the interface as shown in Figure 4.13. This figure shows the surface potential versus the normalized distance along the channel, y/L, for three different conditions of channel length and bias. In a long-channel transistor, the potential is relatively flat over most of the channel length, and the height of the potential barrier is unaffected

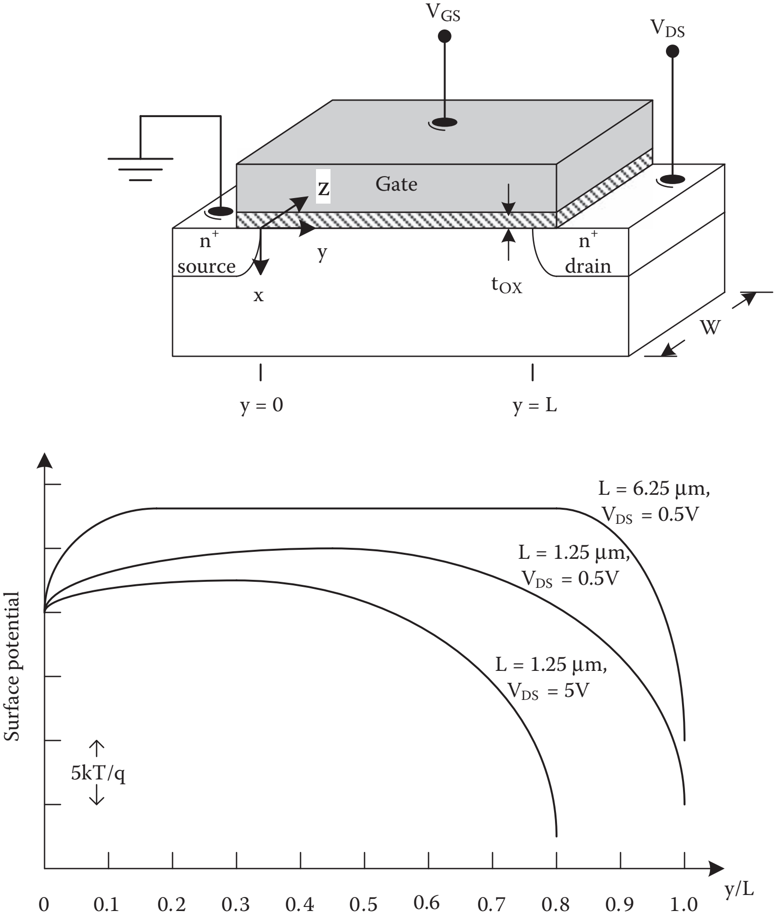


FIGURE 4.13
Surface potential as a function of normalized distance (y/L) for n-MOS transistors with the following: a, L = 6.25 μm and V_{DS} = 0.5V; b, L = 1.25 μm and V_{DS} = 0.5V; and c, L = 1.25 μm and V_{DS} = 5V. (Based on Troutman, R.R., *IEEE Trans. Electron. Dev.*, ED-26, 461, 1979.)

by the drain-to-source bias. This situation is represented by the curve for the case of L = 6.25 μm and V_{DS} = 0.5 V. However, for a transistor with a shorter channel, the potential fails to flatten out so the drain bias affects the height and position of the barrier, as shown by the two curves for the case of L = 1.25 μm. If the drain-to-source bias is increased from 0.5 to 5 V, the barrier height decreases and the position of the peak potential moves toward the source. Therefore, an increase in the drain-to-source bias causes a reduction in the threshold voltage and an increase in the subthreshold current. This contrasts with the case of a long-channel device, for which the subthreshold current is essentially independent of the drain-to-source voltage. Therefore, a strong dependence of the subthreshold current on the drain-to-source bias can be taken as an indicator of short-channel behavior.

4.5.4 Channel Length Modulation

The drain current in a MOSFET saturates at the value of V_{DS} that causes the channel to just pinch off at the drain end. Additional increase in V_{DS} causes the pinch-off point to move into the channel, toward the source, as shown in Figure 4.14.

This increases the drain current by the ratio $L/(L-\Delta L)$. In a long-channel MOSFET, 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 MOSFETs [3]. 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

$$I_D = K\left[\left(V_{GS}-V_T\right)V_{DS}-\frac{V_{DS}^2}{2}\right]\left[1+\lambda V_{DS}\right], \quad V_{GS} \geq V_T \text{ and } \left(V_{GS}-V_T\right) \geq V_{DS}, \quad (4.39)$$

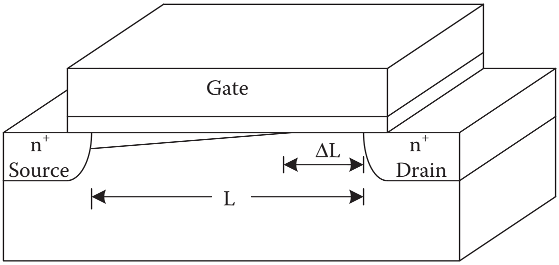


FIGURE 4.14
Channel length modulation in a short-channel MOSFET.

where λ is the empirical *channel length modulation parameter*. Similarly, the equation for the saturated drain current becomes

$$I_D = \frac{K}{2}(V_{GS} - V_T)^2(1 + \lambda V_{DS}); \qquad V_{GS} \geq V_T \text{ and } (V_{GS} - V_T) \leq V_{DS}. \quad (4.40)$$

4.5.5 Field-Dependent Mobility and 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, and this is 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 4.4.2. Also, the magnitude of the saturated drain current is less than that predicted on the basis of the low-field mobility [8].

Empirically, the carrier velocity versus field characteristics can be fit to the following [9,10]:

$$v = \frac{\mu E}{\left[1 + \left(\frac{\mu E}{v_{sat}}\right)^n\right]^{1/n}}, \quad (4.41)$$

where μ is the carrier mobility, E is the electric field intensity, v_{sat} is the carrier saturation velocity, and n is an empirical parameter.

For holes in silicon, $n = 1$ and $v_{satp} = 8 \times 10^6 cm/s$. For electrons in silicon, $n = 2$ and $v_{satn} = 9 \times 10^6 cm/s$. (The saturation velocities in silicon MOSFETs are typically 20% lower than the bulk values.) For approximate hand calculations, we will assume $n \approx 1$ for electrons and holes. Consider first the n-MOS transistor with dimensions and biases as shown in Figure 4.15.

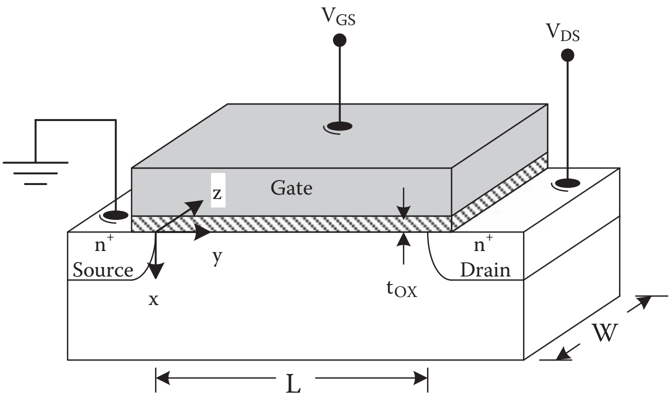


FIGURE 4.15
MOSFET structure for determination of the drain current with velocity saturation.

Our starting point is the assumption that carriers move by drift only and that the drift is only in the y direction parallel to the interface (the *gradual channel approximation*). Based on this assumption, the drain current at a point y along the channel is

$$I_D = -WQ_i(V)\frac{\mu_n dV/dy}{1+(\mu_n/v_{satn})dV/dy}\,. \tag{4.42}$$

Rearranging, we obtain

$$I_D dy = -\left(\mu_n WQ_i(V) + \frac{\mu_n I_D}{v_{satn}}\right)dV\,. \tag{4.43}$$

Integrating both sides, we obtain the drain current for the linear mode of operation,

$$I_D = \frac{-\mu_n(W/L)\int_0^{V_{DS}} Q_i(V)dV}{1+(\mu_n V_{DS}/v_{satn}L)}\,. \tag{4.44}$$

The numerator is the same as the long-channel expression for the drain current, so for linear operation with velocity saturation,

$$I_D = \frac{\mu_n C_{ox}(W/L)\left[(V_{GS}-V_{TN})V_{DS}-V_{DS}^2/2\right]}{1+(\mu_n V_{DS}/v_{satn}L)}\ \text{(linear)}. \tag{4.45}$$

The saturation point occurs at a lower value of drain-to-source voltage with velocity saturation, and this value V_{DSAT} may be found from the solution of

$$dI_D/dV_{DS} = 0\,. \tag{4.46}$$

This yields

$$V_{DSAT} = \frac{2(V_{GS}-V_{TN})}{1+\sqrt{1+2\mu_n(V_{GS}-V_{TN})/(v_{satn}L)}}\,. \tag{4.47}$$

By substituting this into the drain current equation, we can find the saturation current

$$I_D = C_{ox}Wv_{satn}(V_{GS}-V_{TN})\frac{\sqrt{1+2\mu_n(V_{GS}-V_{TN})/(v_{satn}L)}-1}{\sqrt{1+2\mu_n(V_{GS}-V_{TN})/(v_{satn}L)}+1}\ \text{(saturation)}. \tag{4.48}$$

Therefore, when carrier velocity saturation is taken into account the MOS transistor saturates at a lower drain-to-source voltage and with a reduced amount of drain current. This will tend to increase the propagation delays in MOS logic gates. It is also important to note that, in short-channel MOS transistors, 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 so it will not be considered for the purpose of approximate equations used in hand analysis and intuitive device design.

The case of the p-MOS transistor is quite similar, and the approximate drain current equations for both types of transistors are summarized in Tables 4.4 and 4.5.

Example 4.8 Comparison of Long-Channel and Short-Channel Equations for n-MOSFETs

For an n-MOS transistor with $L_N = 2.0\mu m$, $W_N = 4.0\mu m$, and $t_{ox} = 10nm$, compare the drain current characteristics calculated using the long-channel and short-channel equations. Repeat for the case of $L_N = 0.2\mu m$, $W_N = 0.4\mu m$, and $t_{ox} = 5nm$. Assume that $V_{TN} = 0.5V$ in both cases.

Solution: For the $2.0\mu m$ transistor ($L_N = 2.0\mu m$, $W_N = 4.0\mu m$, and $t_{ox} = 10nm$), the long-channel equations are

$$I_D = \begin{cases} 0; & V_{GS} \le 0.5V \\ 400\mu A / V^2 \left[(V_{GS} - 0.5V) V_{DS} - V_{DS}^2 / 2 \right]; & V_{DS} \le (V_{GS} - 0.5V) \\ 400\mu A / V^2 (V_{GS} - 0.5V)^2 / 2; & 0 \le (V_{GS} - 0.5V) \le V_{DS} \end{cases},$$

TABLE 4.4

Drain Current Equations for an n-MOS Transistor, Including the Effect of Carrier Velocity Saturation

Mode	Drain current equation	Voltage conditions
Cutoff	$I_D \approx 0$	$V_{GS} \le V_T$
Linear	$I_D = \dfrac{(C_{ox}\mu_n W / L)\left[(V_{GS} - V_{TN})V_{DS} - V_{DS}^2 / 2\right]}{1 + \mu_n V_{DS} / (v_{satn} L)}$	$V_{DS} \le V_{DSAT}$
Saturation	$I_D = C_{ox} W v_{satn}$ $(V_{GS} - V_{TN}) \dfrac{\sqrt{1 + 2\mu_n (V_{GS} - V_{TN}) / (v_{satn} L)} - 1}{\sqrt{1 + 2\mu_n (V_{GS} - V_{TN}) / (v_{satn} L)} + 1}$	$V_{DS} \ge V_{DSAT}$ $V_{DSAT} = \dfrac{2(V_{GS} - V_{TN})}{1 + \sqrt{1 + 2\mu_n (V_{GS} - V_{TN}) / (v_{satn} L)}}$

Typically, $\mu_n = 580 cm^2 / Vs$ and $v_{satn} = 9 \times 10^6 cm / s$. The drain current is assumed to be entering the drain terminal. Normally, V_{DS}, V_{GS}, and V_{TN} are all positive.

TABLE 4.5

Drain Current Equations for a p-MOS Transistor, Including the Effect of Carrier Velocity Saturation

Mode	Drain current equation	Voltage conditions
Cutoff	$I_D \approx 0$	$V_{GS} \ge V_{TP}$
Linear	$I_D = \dfrac{\left(C_{ox}\mu_p W/L\right)\left[\left(V_{GS}-V_{TP}\right)V_{DS}-V_{DS}^2/2\right]}{1+\mu_p V_{DS}/\left(v_{satp}L\right)}$	$V_{DS} \ge V_{DSAT}$
Saturation	$I_D = -C_{ox}Wv_{satp}$ $\left(V_{GS}-V_{TP}\right)\dfrac{\sqrt{1-2\mu_p\left(V_{GS}-V_{TP}\right)/\left(v_{satp}L\right)}-1}{\sqrt{1-2\mu_p\left(V_{GS}-V_{TP}\right)/\left(v_{satp}L\right)}+1}$	$V_{DS} \le V_{DSAT}$ $V_{DSAT} = \dfrac{2\left(V_{GS}-V_{TP}\right)}{1+\sqrt{1-2\mu_n\left(V_{GS}-V_{TP}\right)/\left(v_{satp}L\right)}}$

Typically, $\mu_p = 230cm^2/Vs$ and $v_{satp} = 8\times 10^6 cm/s$. The drain current is assumed to be leaving the drain terminal. Normally, V_{DS}, V_{GS}, and V_{TP} are all negative.

and the short-channel equations are

$$I_D = \begin{cases} 0; & V_{GS} \le 0.5V \\ \dfrac{400\mu A/V^2\left[\left(V_{GS}-0.5V\right)V_{DS}-V_{DS}^2/2\right]}{1+V_{DS}/3.1V}; & V_{DS} \le V_{DSAT} \\ 1240\mu A/V\left(V_{GS}-0.5V\right)\dfrac{\sqrt{1+\left(V_{GS}-0.5V\right)/1.55V}-1}{\sqrt{1+\left(V_{GS}-0.5V\right)/1.55V}+1}; & V_{DS} \ge V_{DSAT} \end{cases},$$

where

$$V_{DSAT} = \frac{2\left(V_{GS}-0.5V\right)}{1+\sqrt{1+\left(V_{GS}-0.5V\right)/1.55V}}.$$

For the $0.2\mu m$ transistor ($L_N = 0.2\mu m$, $W_N = 0.4\mu m$, and $t_{ox} = 5nm$), the long-channel equations are

$$I_D = \begin{cases} 0; & V_{GS} \le 0.5V \\ 800\mu A/V^2\left[\left(V_{GS}-0.5V\right)V_{DS}-V_{DS}^2/2\right]; & V_{DS} \le \left(V_{GS}-0.5V\right) \\ 800\mu A/V^2\left(V_{GS}-0.5V\right)^2/2; & 0 \le \left(V_{GS}-0.5V\right) \le V_{DS} \end{cases},$$

and the short-channel equations are

$$I_D = \begin{cases} 0; & V_{GS} \le 0.5V \\ \dfrac{800\mu A/V^2\left[\left(V_{GS}-0.5V\right)V_{DS}-V_{DS}^2/2\right]}{1+V_{DS}/0.31V}; & V_{DS} \le V_{DSAT}, \\ 248\mu A/V\left(V_{GS}-0.5V\right)\dfrac{\sqrt{1+\left(V_{GS}-0.5V\right)/0.155V}-1}{\sqrt{1+\left(V_{GS}-0.5V\right)/0.155V}+1}; & V_{DS} \ge V_{DSAT} \end{cases}$$

where

$$V_{DSAT} = \frac{2\left(V_{GS}-0.5V\right)}{1+\sqrt{1+\left(V_{GS}-0.5V\right)/0.155V}}.$$

As shown in Figure 4.16, the long-channel equations provide a reasonably good approximation to the actual characteristics for the $2\mu m$ device, especially at lower values of the gate-to-source bias. Nonetheless, the long-channel equations overestimate the drain current by up to 40% and also predict that the drain current saturates at higher drain-to-source voltages. With the $0.2\mu m$ device, the short-channel equations should be used. The long-channel equations overestimate the saturated drain current by nearly a factor of five for $V_{GS} = 2.0V$. (Generally, the long-channel MOSFET equations provide factor-of-two accuracy down to channel lengths of $\sim 0.6\mu m$.) This example illustrates the necessity of using short-channel equations when analyzing modern MOS transistors, which currently have physical gate lengths down to ~25 nm. Still, it is important to realize that the short-channel equations used for hand analysis fail to capture many of the details of short-channel MOSFET physics so SPICE simulations must be used to accurately predict device behavior and circuit characteristics.

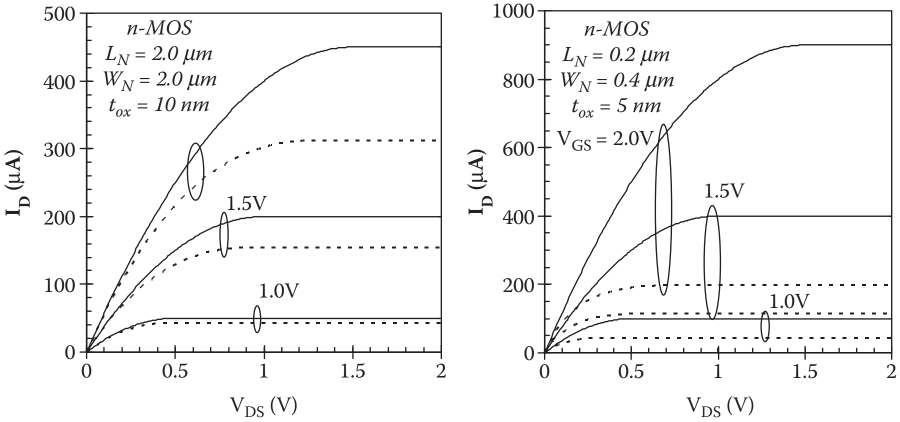


FIGURE 4.16
n-MOS transistor characteristics calculated using long-channel (solid) and short-channel (dashed) equations, for transistors with channel lengths of 2μm and 0.2μm.

Example 4.9 p-MOSFET Long-Channel and Short-Channel Equations

For an p-MOS transistor with $L_P = 2.0\mu m$, $W_P = 4.0\mu m$, and $t_{ox} = 10nm$, compare the drain current characteristics calculated using the long-channel and short-channel equations. Repeat for the case of $L_P = 0.2\mu m$, $W_P = 0.4\mu m$, and $t_{ox} = 5nm$. Assume that $V_{TP} = -0.5V$ in both cases.

Solution: For the $2.0\mu m$ transistor ($L_P = 2.0\mu m$, $W_P = 4.0\mu m$, and $t_{ox} = 10nm$), the long-channel equations are

$$I_D = \begin{cases} 0; & V_{GS} \le 0.5V \\ 160\mu A / V^2 \left[\left(V_{GS} - 0.5V \right) V_{DS} - V_{DS}^2 / 2 \right]; & V_{DS} \le \left(V_{GS} - 0.5V \right) \\ 160\mu A / V^2 \left(V_{GS} - 0.5V \right)^2 / 2; & 0 \le \left(V_{GS} - 0.5V \right) \le V_{DS} \end{cases},$$

and the short-channel equations are

$$I_D = \begin{cases} 0; & V_{GS} \le 0.5V \\ \dfrac{160\mu A / V^2 \left[\left(V_{GS} + 0.5V \right) V_{DS} - V_{DS}^2 / 2 \right]}{1 - V_{DS} / 7.0V}; & V_{DS} \le V_{DSAT} \\ 550\mu A / V \left(V_{GS} + 0.5V \right) \dfrac{\sqrt{1 + \left(V_{GS} + 0.5V \right) / 3.5V} - 1}{\sqrt{1 + \left(V_{GS} - 0.5V \right) / 3.5V} + 1}; & V_{DS} \ge V_{DSAT} \end{cases},$$

where

$$V_{DSAT} = \frac{2\left(V_{GS} - 0.5V \right)}{1 + \sqrt{1 + \left(V_{GS} - 0.5V \right) / 3.5V}}.$$

For the $0.2\mu m$ transistor ($L_N = 0.2\mu m$, $W_N = 0.4\mu m$, and $t_{ox} = 5nm$), the long-channel equations are

$$I_D = \begin{cases} 0; & V_{GS} \le 0.5V \\ 320\mu A / V^2 \left[\left(V_{GS} - 0.5V \right) V_{DS} - V_{DS}^2 / 2 \right]; & V_{DS} \le \left(V_{GS} - 0.5V \right) \\ 320\mu A / V^2 \left(V_{GS} - 0.5V \right)^2 / 2; & 0 \le \left(V_{GS} - 0.5V \right) \le V_{DS} \end{cases},$$

and the short-channel equations are

$$I_D = \begin{cases} 0; & V_{GS} \le 0.5V \\ \dfrac{320\mu A / V^2 \left[\left(V_{GS} - 0.5V \right) V_{DS} - V_{DS}^2 / 2 \right]}{1 + V_{DS} / 0.7V}; & V_{DS} \le V_{DSAT} \\ 220\mu A / V \left(V_{GS} - 0.5V \right) \dfrac{\sqrt{1 + \left(V_{GS} - 0.5V \right) / 0.35V} - 1}{\sqrt{1 + \left(V_{GS} - 0.5V \right) / 0.35V} + 1}; & V_{DS} \ge V_{DSAT} \end{cases},$$

where

$$V_{DSAT} = \frac{2(V_{GS} - 0.5V)}{1 + \sqrt{1 + (V_{GS} - 0.5V)/0.35V}}.$$

As shown by Figure 4.17, it is true here as well that the long-channel equations lead to overestimates of the saturation current and the drain-to-source voltage giving rise to saturation. Because of the smaller low-field mobility of holes compared with electrons, the long-channel MOSFET equations generally provide better accuracy for p-MOS transistors than n-MOS transistors having the same channel length. For p-MOS transistors, factor-of-two accuracy can generally be obtained down to channel lengths of $\sim 0.25\mu m$.

4.5.6 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

$$t_t = \frac{L}{v_{sat}}. \tag{4.49}$$

Therefore, the transit time is directly proportional to the channel length in short-channel MOSFETs.

Example 4.10 Transit Time in a Short-Channel n-MOS Transistor

Estimate the transit time for n-channel MOSFETs with a 65 nm channel length in a 1.5 V CMOS circuit.

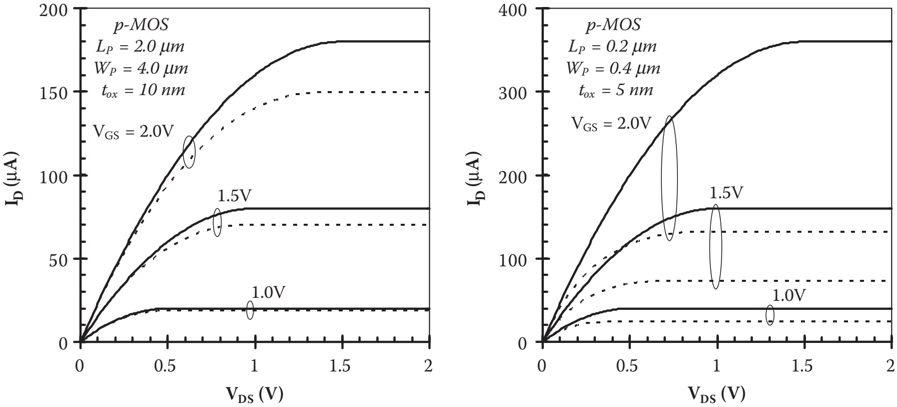


FIGURE 4.17
p-MOS transistor characteristics calculated using long-channel (solid) and short-channel (dashed) equations, for transistors with channel lengths of 2μm and 0.2μm.

Solution: The average channel field for a saturated MOSFET is approximately

$$E \approx \frac{1.5V}{65 \times 10^{-7} cm} = 2.3 \times 10^{5} V / cm \cdot$$

Therefore, it is appropriate to assume that the electrons drift at their saturated velocity in the channel. The transit time is

$$t_t = \frac{L}{v_{sat}} = \frac{65 \times 10^{-7} cm}{10^{7} cm / s} = 0.65 ps \cdot$$

The quasi-static approximation is applicable if the circuit propagation delay is greater than this transit time.

4.6 MOSFET Design

The design rules for MOSFETs are based on the minimum dimensions, separations, and surrounds given in Table 4.6.

The basic design rules for a MOSFET are illustrated in Figure 4.18 for the case of an n-MOS transistor. The channel is formed when the polysilicon wire overlaps the active region. The minimum width for this polysilicon wire, and therefore the "printed gate length L," is 2X (rule *L1*), and the polysilicon wire must extend beyond the active region by at least 1X on either side (rule *L2*). The nselect implant (not shown in these simplified layout drawings) must extend 1X (rule *S6*) beyond the edges of the active region. The contact windows are always 2X square (rule *L3*), but multiple contact openings are used

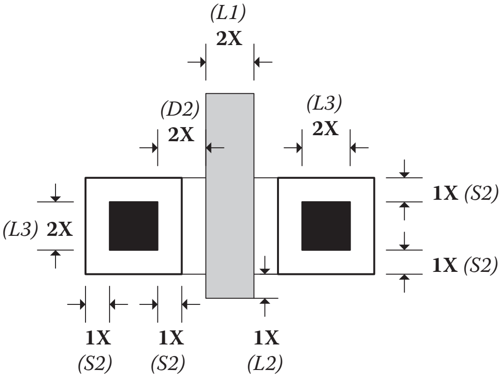


FIGURE 4.18
n-MOS transistor design rules.

TABLE 4.6

Scalable Layout Design Rules Adopted for Use in This Book

Rule	Description	Value
Minimum dimensions		
L1	Gate length/polysilicon width	2X
L2	Extension of polysilicon gate beyond active region	1X
L3	Width of contact window	2X
L4	Width of active region	3X
L5	Width of implanted region	3X
L6	Width of metal 1	3X
L7	Width of metal 2	3X
Minimum separations		
D1	Spacing between polysilicon gates/ wires	2X
D2	Spacing between polysilicon gate and S/D contact window	2X
D3	Spacing between contacts	2X
D4	Spacing between active regions	3X
D5	Spacing between implanted regions of same type	3X
D6	Spacing between metal 1 wires	3X
D7	Spacing between metal 2 wires	4X
D8	Spacing between implanted regions of opposite type	5X
Minimum surrounds		
S1	Active region surrounding contact window	1X
S2	Metal 1 surrounding contact window	1X
S3	Metal 2 surrounding contact window	1X
S4	Polysilicon surrounding contact window	1X
S4	nselect or pselect surrounding contact window	1X
S6	nselect or pselect surrounding active region	1X
S7	n-Well surrounding p-MOS active region	5X

if additional contact area is needed. The metal placed over a contact opening must extend beyond the window edges by 1X in all directions (rule *S2*), and this metal must be spaced by at least 1X from the channel region. 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.

In the case of a p-MOS transistor, an n-well surrounds the device as shown in Figure 4.19. The n-well must extend at least 5X out from the active region in all directions (rule *S7*). These scalable design rules apply for an n-well process, in which the n-MOS transistor is fabricated directly in the p-type substrate. In the case of a twin-well process, a p-type well is created for the n-MOS device, so there are additional rules associated with the p-well dimensions. If absolute (rather than scalable) rules are used, they will apply to the same dimensions, separations, and extensions described above, but they will be in units of length rather than multiples of X.

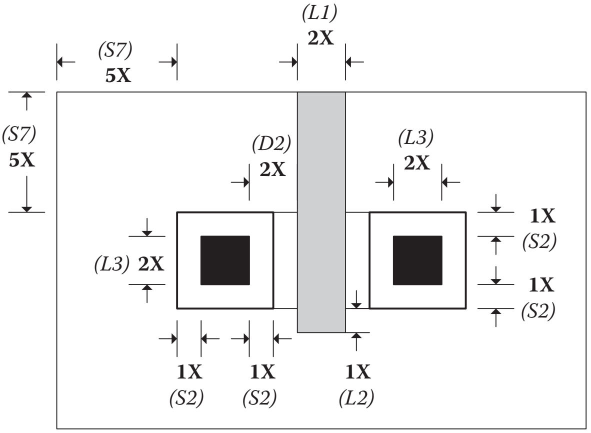


FIGURE 4.19
p-MOS transistor design rules.

Often it is necessary to connect two or more MOSFETs in series as shown in Figure 4.20 for the case of two n-MOS transistors. In such a situation, it is not necessary to form source and drain contact regions between the series connected transistors; instead, the common nselect region between the two channels forms the drain of one transistor and the source of the other. The minimum separation between the two polysilicon gates is 2X (rule *D1*).

In many situations, it is also necessary to place two MOS transistors in parallel, to create the desired logic function or perhaps just to increase the overall current drive capability, as shown in Figure 4.21. These two transistors share a single source region. Multiple contacts have been used for both drains as well as the common source. These contact areas are 2X in width (rule *L3*) and are spaced by 2X (rule *D3*).

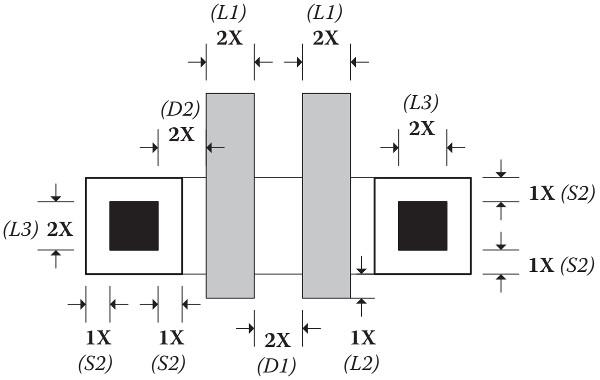


FIGURE 4.20
Design rules for series-connected n-MOS transistors.

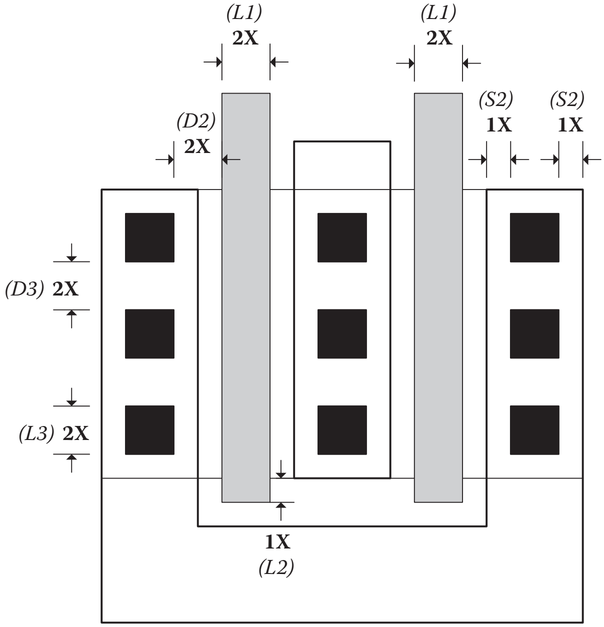


FIGURE 4.21
Design rules for parallel-connected n-MOS transistors.

Example 4.11 Layout Area for n-MOS Transistors

Estimate the layout area, in terms of X^2, for an n-MOS transistor with a $K = 1mA/V^2$. Repeat for an n-MOSFET with double the device transconductance parameter. Account for the fact that nselect regions of adjacent n-MOS transistors must be separated by at least 3X and assume that $t_{ox} = 9nm$.

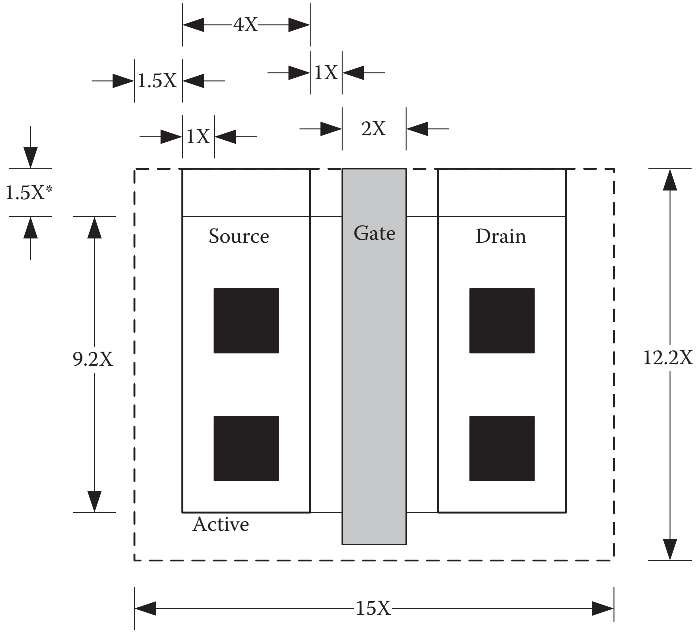
Solution: The process transconductance parameter is

$$k'_N = \frac{\mu_n \varepsilon_{ox}}{t_{ox}} = \frac{\left(580cm^2V^{-1}s^{-1}\right)(3.9)\left(8.85\times10^{-14}F/cm\right)}{9\times10^{-7}cm} = 220\mu A/V^2 \cdot$$

The required aspect ratio is

$$\frac{W_N}{L_N} = \frac{K_N}{k'_N} = \frac{1000\mu A/V^2}{220\mu A/V^2} = 4.6\ ,$$

and, assuming the minimum gate length of 2X is used, the gate width should be 9.2X. The transistor layout using this width is shown in Figure 4.22. The contact windows are all made 2X square and must be separated by 2X. The number of contact windows is determined by the maximum number that will fit using these rules and is therefore two in this device. If we attribute one-half of the separation between the nselect regions to this transistor, then the overall area



* Active areas must be spaced by at least 3X.

FIGURE 4.22
Example layout of a n-MOSFET with $K = 1mA / V^2$.

is $A_{N1000} = (12.2X)(15X) \approx 183X^2$. Using a minimum gate length of 0.6 μm, $X = 0.3\mu m$, and $A_{N1000} \approx 16.5\mu m^2$.

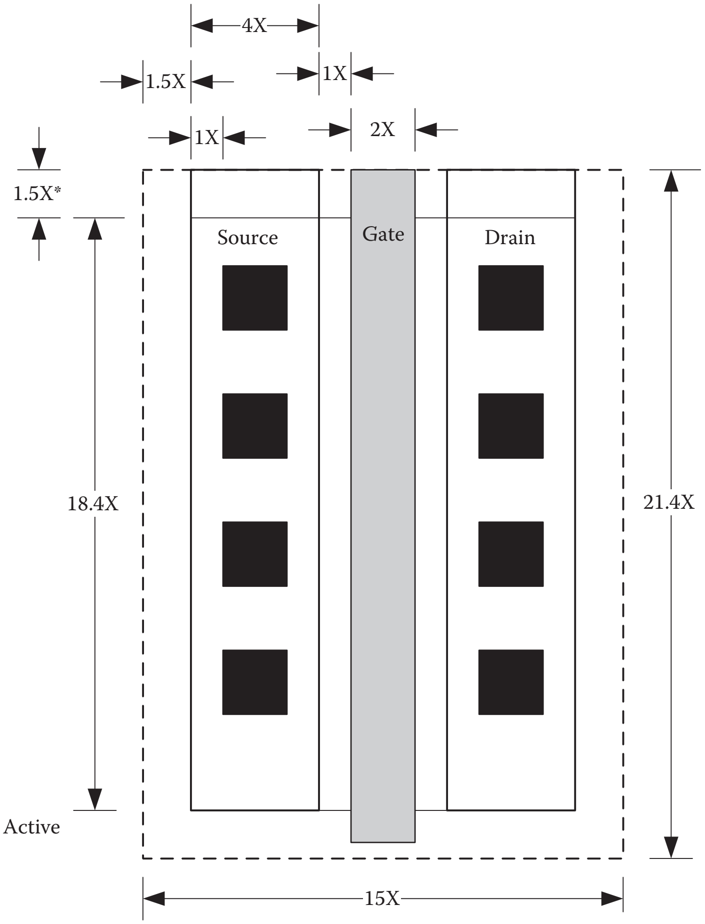
For $K = 2mA / V^2$, the required aspect ratio is doubled to 9.2 and the gate width should be 18.4X. The transistor layout using the increased width is shown in Figure 4.23. Here four contact windows should be used for the source and drain. The overall area is $A_{N2000} = (21.4X)(15X) \approx 321X^2$. Using a minimum gate length of $2X = 0.6\mu m$, $A_{N2000} \approx 28.9\mu m^2$; in other words, doubling the device transconductance parameter did not double the required layout area. Here, the 100% increase in K required a 75% increase in the device area, but this ratio will generally depend on the absolute device widths attributable to the constant (overhead) contributions.

Example 4.12 Layout Area for p-MOS Transistors

Estimate the layout area, in terms of X^2, for a p-MOS transistor with a $K = 1mA / V^2$ Account for the n-well with its minimum extension of 5X in all directions beyond the pselect region and assume that $t_{ox} = 9nm$.

Solution: The process transconductance parameter is

$$k_P' = \frac{\mu_P \varepsilon_{ox}}{t_{ox}} = \frac{(230cm^2V^{-1}s^{-1})(3.9)(8.85 \times 10^{-14} F / cm)}{9 \times 10^{-7} cm} = 88\mu A / V^2 .$$



* Active areas must be spaced by at least 3X.

FIGURE 4.23
Example layout of a n-MOSFET with $K = 2mA / V^2$.

The required aspect ratio is

$$\frac{W_P}{L_P} = \frac{K_P}{k'_P} = \frac{1000\mu A / V^2}{88\mu A / V^2} = 11.4 ,$$

and, assuming the minimum gate length of 2X is used, the gate width should be 22.8X. The transistor layout using this width is shown in Figure 4.24. Five contact windows may be used each for the source and drain. If the n-well surround (5X in all directions) is included, the overall area is $A_{P1000} = (22X)(32.8X) \approx 722X^2$. Using a minimum gate length of $2X = 0.6\mu m$, $A_{P1000} \approx 65\mu m^2$. The actual area associated with the transistor will vary depending on whether the adjacent transistors are n-MOS or p-MOS devices. A neighboring p-MOS device can share the n-well, thereby reducing the effective area per device. Either way, a p-MOS device consumes more area than an n-MOSFET having the same device transconductance value.

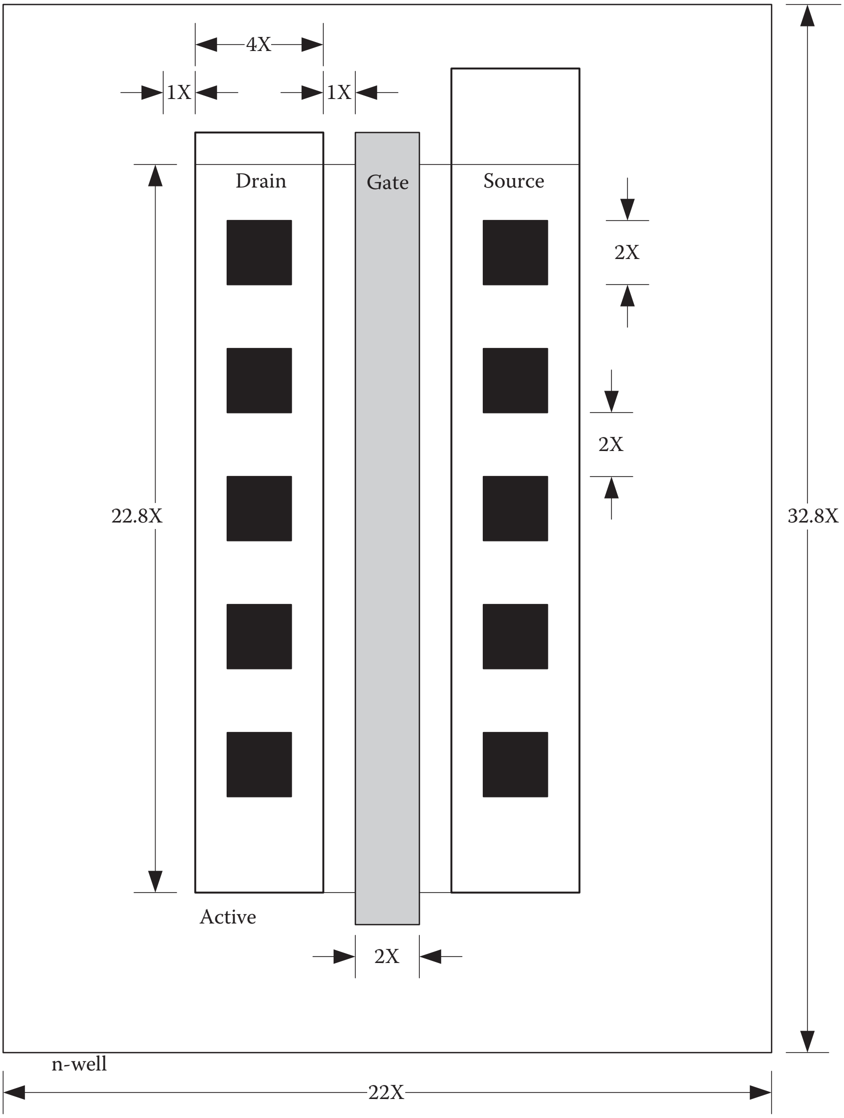


FIGURE 4.24
Example layout of a p-MOSFET with $K = 1mA/V^2$.

4.7 MOSFET Capacitances

For practical CMOS digital circuits, the propagation delays exceed the carrier transit times by a factor of 10; therefore, the parasitic capacitances limit the switching speed. Most of this capacitance is associated with the MOS transistors themselves, both the driving devices and the load devices, although occasionally the interconnect capacitance is important as well.

There are two important types of capacitances in a MOSFET: (1) the capacitances associated with the gate oxide, and (2) those associated with the depletion regions of p-n junctions. The oxide capacitances appear between the gate and the source, the gate and the drain, and the gate and the body. The p-n junction capacitances appear between the S/D implanted regions and the body. Figure 4.25 shows the circuit diagram of an n-MOS transistor with its five important capacitances: C_{gs}, C_{gd}, and C_{gb} are oxide capacitances, whereas C_{db} and C_{sb} are p-n junction capacitances.

A third type of parasitic capacitance is introduced as a result of the fringing electric fields between the gate and the S/D contacts. Although this contribution may be neglected in long-channel devices, it plays an increasingly important role in deeply scaled MOS devices.

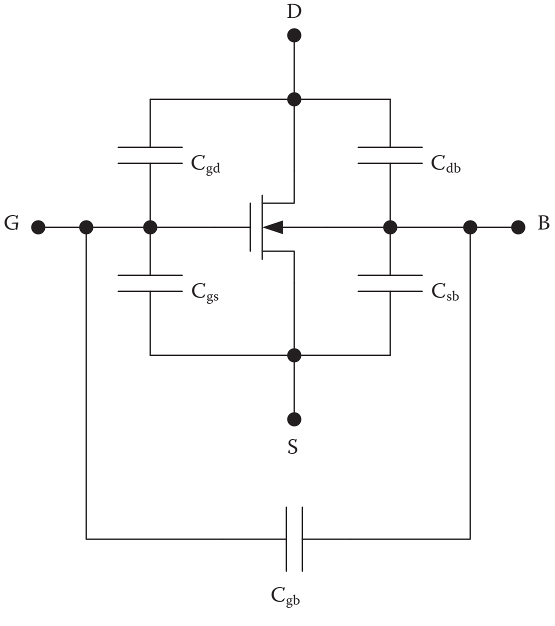


FIGURE 4.25
An n-MOS transistor with the important device capacitances.

4.7.1 Oxide Capacitances

The oxide capacitances in a MOSFET are distributed throughout device and depend on the biasing, making their exact determination quite complex. Fortunately, some simple approximations may be made based on the physical pictures of Figure 4.26a–c for the purpose of hand performance calculations.

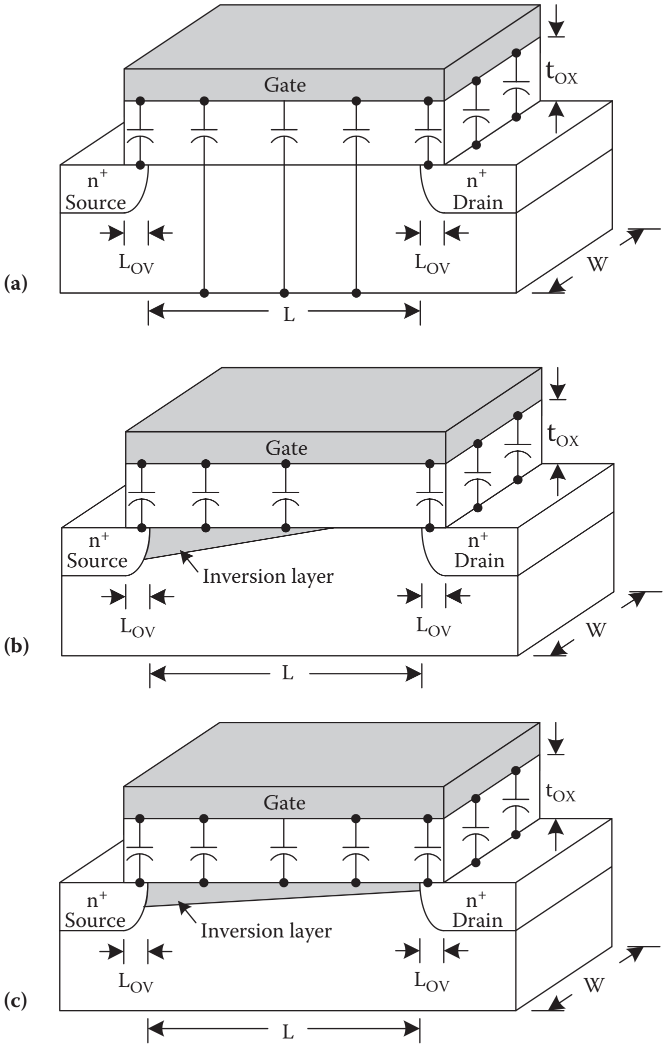


FIGURE 4.26
Oxide capacitances in an n-MOS transistor for (a) cutoff, (b) saturated, and (c) linear operation.

These illustrations show the oxide contributions to the device capacitances for an n-MOS transistor in the cutoff, saturation, and linear modes of operation, respectively (Table 4.7).

In the case of cutoff operation (Figure 4.26a), there is no inversion layer, so the parallel plate capacitance WLC_{ox}, of the gate oxide appears between the gate and the p-type body. The contributions to C_{gs} and C_{gd} are attributable to the overlap of the gate with the S/D implantations. If we assume that the device structure is symmetric, with equal overlaps at the source and drain, then $C_{gs}(cutoff) \approx C_{gd}(cutoff) \approx WL_{OV}C_{ox}$.

For saturation operation (Figure 4.26b), part of the parallel plate capacitance appears between the gate and source. The exact value of C_{gs} depends on the details of the device geometry, material parameters, and bias point; however the usual rule of thumb is to assume $C_{gs}(sat) \approx 2WLC_{ox}/3 + WL_{OV}C_{ox}$. Because the channel pinches off at the drain end, the parallel plate capacitance does not contribute to the gate-to-drain capacitance and $C_{gd}(sat) \approx WL_{OV}C_{ox}$. The electrons in the channel shield the body from the gate so that $C_{gb}(sat) \approx 0$.

In the linear region of operation (Figure 4.26c), the inversion layer extends from the source to the drain, and the parallel plate capacitance of the oxide splits roughly evenly between the source and the drain: $C_{gs}(linear) \approx C_{gd}(linear) \approx WLC_{ox}/2 + WL_{OV}C_{ox}$. Here again the shielding behavior of the channel carriers results in $C_{gb}(linear) \approx 0$.

The simple approximations given here do not account for the detailed variation of the capacitances with the applied voltages. Also, for a MOSFET in a logic gate circuit, the Miller effect may increase the effective value of a feedback capacitance, such as C_{gd}. For hand calculations, however, it is often adequate to assume that the three capacitances act in parallel, between the gate and AC ground, with a total worst-case value of $C_g \approx WLC_{ox} + 2WL_{OV}C_{ox}$.

The case of a p-MOS transistor is fundamentally similar. However, the drain/source implantations use a separate implantation step, using a different dopant (boron) and different implantation conditions. Therefore, the S/D overlaps will not be the same as for the n-MOS devices.

TABLE 4.7

Oxide Capacitances in a MOSFET

Mode	C_{gs}	C_{gb}	C_{gd}
Cutoff	$WL_{OV}C_{ox}$	WLC_{ox}	$WL_{OV}C_{ox}$
Saturation	$2WLC_{ox}/3 + WL_{OV}C_{ox}$	0	$WL_{OV}C_{ox}$
Linear	$WLC_{ox}/2 + WL_{OV}C_{ox}$	0	$WLC_{ox}/2 + WL_{OV}C_{ox}$

4.7.2 p-n Junction Capacitances

As discussed in Chapter 3, any p-n junction exhibits a voltage-dependent capacitance associated with its depletion layer. In a MOSFET, the source and drain regions form p-n junctions with the body region of opposite conductivity type (either the substrate or n-well). This results in parasitic capacitances C_{sb} and C_{gb} that can greatly influence switching speed.

As with the oxide contributions, these are distributed capacitances associated with complex device geometries so that their exact calculation is complex. However, we can make some useful estimates based on the simple geometry shown in Figure 4.27. Here the source and drain regions are assumed to be box shaped with a junction depth x_j. The width of the device is W, and the lengths of the source and drain are L_S and L_D, respectively. C_{bot} and C_{sw} represent the depletion layer capacitances per unit area for the bottom and sidewall junctions, respectively; C_{sw} will be considerably higher than C_{bot} if a *channel stopper* implant is used. This channel stopper makes the doping heavier than the surrounding body (p^+ for an n-MOS transistor) to prevent the formation of a parasitic channel between devices. As a consequence, the depletion width is reduced and the junction capacitance is increased.

The total junction capacitance acting between the source and the body involves the bottom and three sidewalls of the implanted source region. The capacitance of the sidewall facing the channel region can often be neglected. If the depletion layer capacitances per unit area are C_{bot} and C_{sw} for the bottom and sidewalls, respectively, then the source capacitance is

$$C_{sb} \approx WL_S C_{bot} + x_j \left(2L_S + W\right) C_{sw} . \tag{4.50}$$

Similarly, the total depletion layer capacitance connected to the drain is

$$C_{db} \approx WL_D C_{bot} + x_j \left(2L_D + W\right) C_{sw} . \tag{4.51}$$

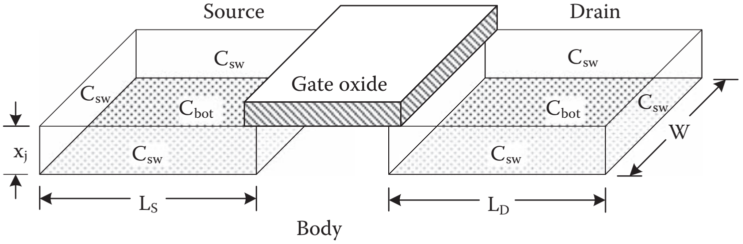


FIGURE 4.27
Junction capacitances associated with the S/D implanted regions of a MOSFET.

Generally, the depletion capacitance for a p-n junction depends on the voltage bias. For hand calculations, the usual practice is to use an effective value of capacitance, averaged over the range of applied voltage. Thus, if the junction voltage varies from an initial value V_1 to a final value of V_2, the effective junction capacitance is

$$C_J = \frac{C_{J0}V_{bi}}{(V_1 - V_2)(1-m)}\left[\left(1-\frac{V_2}{V_{bi}}\right)^{1-m} - \left(1-\frac{V_1}{V_{bi}}\right)^{1-m}\right], \tag{4.52}$$

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 and varies from ⅓ for a linearly graded junction to ½ for an abrupt junction.

Figure 4.28 shows a simplified model for the capacitances in an n-MOS transistor with the source and body grounded. Here the three oxide-related capacitances act approximately in parallel. (Note however that, if the drain is not at AC ground, the Miller effect will alter the effective value of C_{gd}.) The source-to-body capacitance C_{sb} is unimportant if the source and body are both grounded.

Example 4.13 MOSFET Capacitances

Estimate the capacitances associated with the n-MOS transistor shown in Figure 4.29, assuming that the source and body are grounded, V_{GS} = 5V, and the drain makes a transition from 5 to ~0 V. The dimensions of the device are as follows: W = 2 μm, t_{ox} = 25 nm, L = 0.5 μm, $L_D = L_S$ = 2 μm, x_j = 0.2 μm, and L_{OV} = 0.1

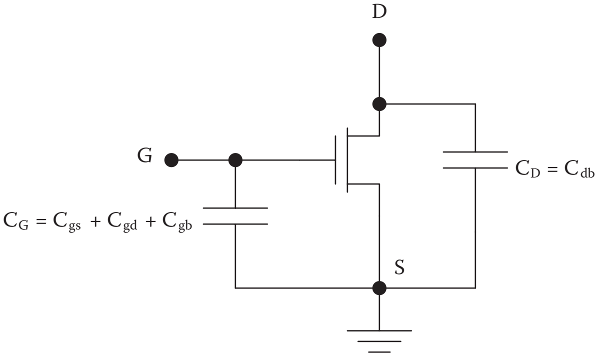


FIGURE 4.28
Simplified model for the capacitances in an n-MOS transistor.

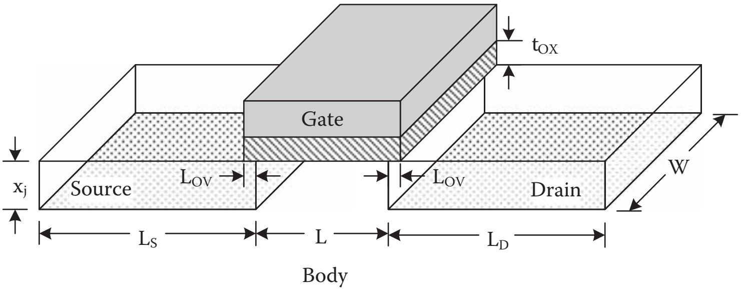


FIGURE 4.29
Example n-MOS transistor for the calculation of device capacitances.

μm. Assume abrupt source and drain junctions with $N_d = 10^{19}$ cm^{-3}. The substrate doping concentration is $N_a = 10^{16}$ cm^{-3}, whereas the sidewall (channel stopper) doping is $N_a = 10^{17}$ cm^{-3}.

Solution: The oxide capacitance per unit area is

$$C_{ox} = \frac{\varepsilon_{ox}}{t_{ox}} = \frac{(3.9)(8.85\times10^{-14}F/cm)}{25\times10^{-7}cm} = 1.38\times10^{-7}F/cm^2 \cdot$$

The worst-case oxide capacitance between the gate and AC ground is

$$\begin{aligned} C_G &\approx WLC_{ox} + 2WL_{OV}C_{ox} \\ &= (2\times10^{-4}cm)(0.5\times10^{-4}cm)(1.38\times10^{-7}F/cm^2) \\ &\quad +2(2\times10^{-4}cm)(0.1\times10^{-4}cm)(1.38\times10^{-7}F/cm^2) \\ &= 1.38fF + 0.55fF = 1.93fF. \end{aligned}$$

For the p-n junctions at the bottom of the source and drain diffusions,

$$V_{bi} = \frac{kT}{q}\ln\left(\frac{N_aN_d}{n_i^2}\right) = (0.026V)\ln\left(\frac{(10^{16}cm^{-3})(10^{19}cm^{-3})}{(1.45\times10^{10}cm^{-3})^2}\right) = 0.88V\ ,$$

$$W = \sqrt{\frac{2\varepsilon_s V_{bi}}{qN_a}} = \sqrt{\frac{2(11.9)(8.85\times10^{-14}F/cm)(0.88V)}{(1.602\times10^{-19}C)(10^{16}cm^{-3})}} = 0.34\times10^{-4}cm = 0.34\mu m\ '$$

and the zero-bias depletion capacitance per unit area is

$$C_{bot0} = \frac{(11.9)(8.85\times10^{-14}F/cm)}{0.34\times10^{-4}cm} = 3.1\times10^{-8}F/cm^2 \cdot$$

For the p-n junctions at the sidewalls,

$$V_{bi}=\frac{kT}{q}\ln\left(\frac{N_aN_d}{n_i^2}\right)=(0.026V)\ln\left(\frac{\left(10^{17}cm^{-3}\right)\left(10^{19}cm^{-3}\right)}{\left(1.45\times10^{10}cm^{-3}\right)^2}\right)=0.94V,$$

$$W=\sqrt{\frac{2\varepsilon_sV_{bi}}{qN_a}}=\sqrt{\frac{2(11.9)\left(8.85\times10^{-14}F/cm\right)(0.94V)}{\left(1.602\times10^{-19}C\right)\left(10^{17}cm^{-3}\right)}}=0.111\times10^{-4}cm=0.111\mu m,$$

and with zero bias,

$$C_{sw0}=\frac{(11.9)\left(8.85\times10^{-14}F/cm\right)}{0.111\times10^{-4}cm}=9.5\times10^{-8}F/cm^2.$$

The source-to-body capacitance need not be considered because both the body and source are grounded. The drain makes a transition from 5 to ~0 V; with the body grounded, V_{db} transitions from −5 to 0 V. (The minus sign accounts for the reverse bias on the drain-to-body p-n junction.) The effective junction capacitances are

$$C_{bot}=\frac{\left(3.1\times10^{-8}F/cm^2\right)(0.88V)}{(-5V)(0.5)}\left[1-\left(1-\frac{-5V}{0.88V}\right)^{0.5}\right]=1.73\times10^{-8}F/cm^2$$

and

$$C_{sw}=\frac{\left(9.5\times10^{-8}F/cm^2\right)(0.94V)}{(-5V)(0.5)}\left[1-\left(1-\frac{-5V}{0.94V}\right)^{0.5}\right]=5.4\times10^{-8}F/cm^2.$$

The effective capacitance acting between the drain and ground is thus

$$\begin{aligned}C_D&\approx WL_DC_{bot}+x_j\left(2L_D+W\right)C_{sw}\\&=\left(2\times10^{-4}cm\right)\left(2\times10^{-4}cm\right)\left(1.73\times10^{-8}F/cm^2\right)\\&\quad+\left(0.2\times10^{-4}cm\right)\left(4\times10^{-4}cm+2\times10^{-4}cm\right)\left(5.4\times10^{-8}F/cm^2\right)\\&=0.70fF+0.65fF=1.35fF.\end{aligned}$$

The resulting model is shown in Figure 4.30. In this example, C_G and C_D are comparable in value. Despite this, the input capacitance C_G is usually far more important in determining circuit speed. Consider an inverter made from the example n-MOS transistor. Most of the capacitance loading the output node (drain) is associated with the gate capacitances of the N fan-out gates. Only under unloaded or lightly loaded (small N) conditions will the output capacitance C_D be important in speed calculations.

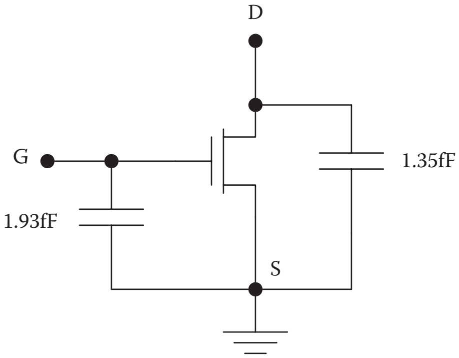


FIGURE 4.30
Model of example n-MOS transistor with parasitic capacitances.

4.7.3 The Miller Effect

For an n-MOS transistor used in a inverter circuit with the source grounded as shown in Figure 4.31, the displacement current in the feedback capacitance C_{gd} depends on both the input and output voltage waveforms. If, for example, the input voltage makes a transition from zero to V_{DD} whereas the output makes a transition from V_{DD} to zero, the voltage v_C across the feedback capacitor makes a transition from $-V_{DD}$ to $+V_{DD}$; in other words, C_{gd} experiences a voltage swing of $2V_{DD}$. If we model the effect of this capacitance at the input with an effective capacitance connected to ground, then its value must be $2C_{gd}$ to correctly account for the doubled average displacement current in the actual feedback capacitor. Similarly, to model the effect of the feedback capacitance at the output with a grounded capacitance, we

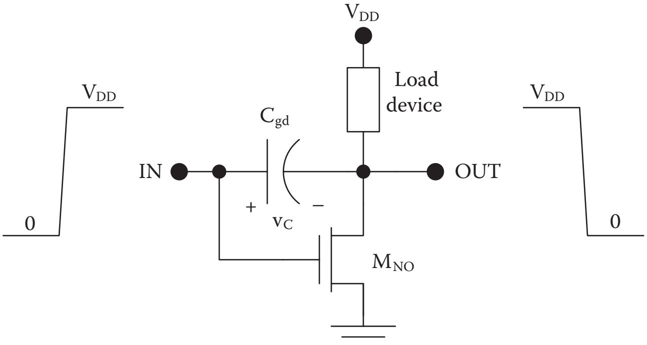


FIGURE 4.31
MOS inverter with a feedback capacitance for consideration of the Miller effect.

must use an effective value of $2C_{gd}$. Figure 4.32 shows the MOS inverter with the ground-connected capacitances. Here C_{Mi} and C_{Mo} are referred to as the *input* and *output Miller capacitances*, respectively, and this overall behavior is termed the *Miller effect*.

Here we should emphasize two important points regarding the Miller effect. (1) In a MOS digital circuit, the feedback capacitance is effectively doubled with regard to its effect on the circuit speed. (2) Use of the simple model presented here, with $C_{Mi} = 2C_f$, allows accurate prediction of only the *average* displacement current in the feedback capacitance for hand calculations. The instantaneous displacement current in the actual circuit might depart considerably from the value predicted using the Miller input capacitance, because in general $C_{Mi} = C_f(1 - A)$, where A is the voltage gain for the circuit. To obtain $C_{Mi} = 2C_f$, we have assumed the average voltage gain A = –1 for an inverter, but digital inverter circuits are highly nonlinear.

4.8 MOSFET Constant-Field Scaling

The steady improvements in the performance and density of CMOS digital integrated circuits over the past few decades have been achieved by reduction, or *scaling*, of the device dimensions. Ideally, the scaling of device dimensions should be accompanied by voltage reductions by the same

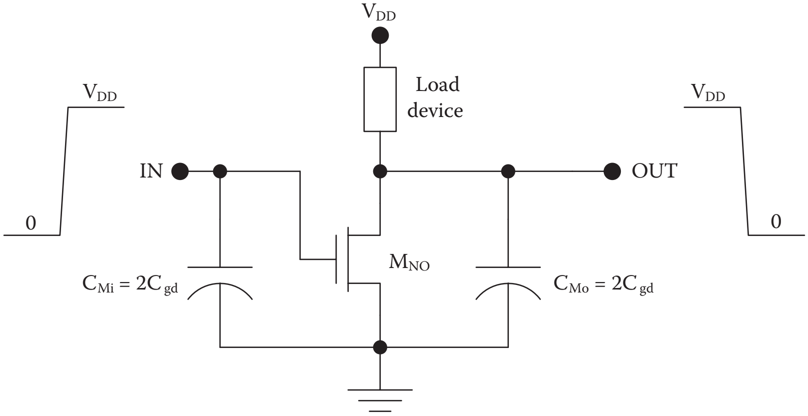


FIGURE 4.32
MOS inverter with shown with effective capacitances acting between the input and ground (the input Miller capacitance C_{Mi}) and between the output and ground (the output Miller capacitance C_{Mo}).

factor so that the electric field intensities remain unchanged. This approach helps to avoid high-field effects such as increased leakage currents, device degradation, and breakdown. Thus, if the device dimensions W, L, t_{ox}, x_j, L_D, L_S, and L_{OV} are all scaled by a factor $1/\kappa$, where $\kappa > 1$, then the voltages V_{DD} and V_T should also be scaled by the same factor $1/\kappa$. To scale the junction depletion widths by this same factor, the doping concentrations must be scaled by approximately κ. Table 4.8 summarizes the ideal case of constant-field scaling. In Chapter 6, we will revisit the issue of scaling, with a discussion of other scaling approaches and how they impact *circuit* performance.

4.9 SPICE MOSFET Models

The complex physical behavior of modern short-channel MOS transistors mandates use of the computer tool SPICE for accurate circuit performance predictions. All SPICE MOSFET models are based on the circuit diagram of Figure 4.33. They differ primarily in the complexity of the model for the dependent current source. The level 1 model is the simplest, allowing short computation times while providing reasonable accuracy for most purposes. The level 2 model is more accurate, accounting for the bias dependent carrier mobility and carrier velocity saturation. It also provides a more accurate model for the channel length modulation. The level 3 model was developed to provide accuracy similar to the level 2 model but with shorter computational times. However, none of these models achieves the required level of accuracy for design and analysis of modern CMOS digital circuits,

TABLE 4.8

Constant-Field Scaling of MOSFET Devices

Device parameters	Scale multiplier
L, W, t_{ox}, x_j	$1/\kappa$
V_{DD}, V_T	$1/\kappa$
N_a, N_d	κ
C_{ox}	κ
C_G	$1/\kappa$
k′	κ
K	κ
I_D	$1/\kappa$
Device area	$1/\kappa^2$

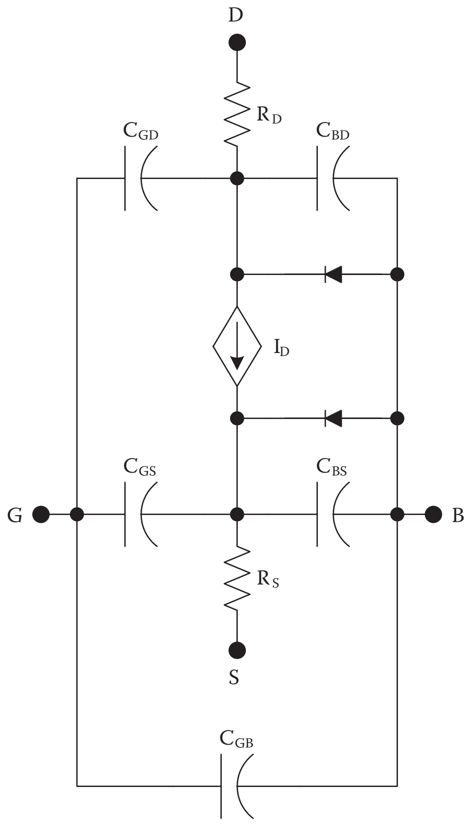


FIGURE 4.33
SPICE model for the MOSFET.

so they have been supplanted by the Berkeley short-channel IGFET* model (BSIM) [11–16].

Here the level 1 model will first be described because it bears a close resemblance to the equations developed for hand calculations and because it serves as a good starting point for work with SPICE. The BSIM will also be described because of its prevalence in industry. This model is not used in a course on digital integrated circuits because its complexity actually obscures some of the relationships between device parameters and circuit performance. However, the reader should be aware that a transition to VLSI design with deep-submicrometer MOS transistors will require use of a BSIM to obtain accurate performance predictions.

* Insulated gate field effect transistor (IGFET) is another name for the MOSFET.

4.9.1 MOSFET Level 1 Model

The level 1 MOSFET model (Table 4.9) is based on the circuit diagram of Figure 4.33. Here R_D and R_S are the parasitic resistances in the drain and source, respectively. C_{GS}, C_{GD}, and C_{GB} are the gate-to-source, gate-to-drain, and gate-to-body capacitances, respectively. C_{BS} and C_{BD} are the body-to-source and body-to-drain capacitances, respectively. The diodes are parasitic p-n junctions between the S/D regions and the body; they are reverse biased under normal operating conditions.

TABLE 4.9

MOSFET Level 1 SPICE Parameters

Symbol	Spice name	Description	Units	Default	Typical
k'	KP	Process transconductance parameter	A/V^2	2E–5	1E–4
V_{TO}	VTO	Threshold voltage with $V_{BS} = 0$	V	1.0	0.5
γ	GAMMA	Body effect coefficient	$V^{1/2}$	0	0.30
$\|2\phi_F\|$	PHI	Surface inversion potential	V	0.6	0.7
λ	LAMBDA	Channel-length modulation parameter	V^{-1}	0	0.02
t_{ox}	TOX	Oxide thickness	M	1E–7	2E–8
N_a	NSUB	Substrate doping	cm^{-3}	0	1E15
μ_0	UO	Carrier mobility	cm^2/Vs	600	580
I_s	IS	S/D junction reverse saturation current	A	1E–14	1E–15
J_s	JS	S/D junction reverse saturation current density	A/m^2	0	1E–8
V_{bi}	PB	S/D junction built-in voltage	V	0.8	0.75
C_{j0}	CJ	S/D junction capacitance per unit area at zero bias	F/m^2	0	2E–4
M_j	MJ	S/D junction grading coefficient		0.5	0.5
C_{j0sw}	CJSW	S/D sidewall junction capacitance per unit periphery	F/m	0	1E–9
M_{jsw}	MJSW	S/D sidewall junction grading coefficient		0.33	0.33
C_{GSO}	CGSO	Gate-source overlap capacitance per unit gate width	F/m	0	4E–10
C_{GDO}	CGDO	Gate-drain capacitance per unit gate width	F/m	0	4E–10
C_{GBO}	CGBO	Gate-body overlap capacitance per unit gate width	F/m	0	2E–10
R_D	RD	Drain series resistance	Ω	0	10
R_S	RS	Source series resistance	Ω	0	10

For the calculation of the drain current, the level 1 model accounts for channel length modulation and the modification of the threshold voltage by the body effect, but carrier velocity saturation is not included. The level 1 equations for the threshold voltage and drain current are therefore as follows:

$$V_T = VTO + GAMMA\left(\sqrt{PHI + V_{SB}} - \sqrt{PHI}\right); \tag{4.53}$$

$$I_D = \frac{KP}{2}\frac{width}{length}\left(V_{GS} - V_T\right)^2\left(1 + V_{DS}LAMBDA\right), \text{(saturation)}; \tag{4.54}$$

and

$$I_D = \frac{KP}{2}\frac{width}{length}\left[\left(V_{GS} - V_T\right)V_{DS} - \frac{V_{DS}^2}{2}\right]\left(1 + V_{DS}LAMBDA\right), \text{(linear)}, \tag{4.55}$$

where V_{GS} is the gate-to-source voltage, V_{DS} is the drain-to-source voltage, V_{SB} is the source-to-body voltage, I_D is the drain current source, VTO is the threshold voltage with zero body to source bias, PHI is $2\phi_F$, width is the gate width, length is the gate length, KP is the process transconductance parameter, and LAMBDA is the channel length modulation parameter. KP, LAMBDA, VTO, and PHI are SPICE model parameters, but width and length are not; instead, these are associated with particular devices.

There are redundant SPICE model parameters that allow alternative means to specify a device model. For example, it is possible to enter the oxide thickness (TOX) and carrier mobility (UO) rather than the process transconductance parameter, which will then be calculated from

$$KP = \frac{UO}{TOX}\varepsilon_{ox}. \tag{4.56}$$

If UO, TOX, and KP are all specified in the model, then the highest level parameter (KP) will override the others. In similar manner, GAMMA may be specified directly in the model or calculated from the values of NSUB and TOX:

$$GAMMA = \frac{\sqrt{2q\varepsilon_{Si}NSUB}}{\varepsilon_{ox}/TOX}. \tag{4.57}$$

The voltage-dependent device capacitances are calculated by including the parallel plate gate oxide capacitance, the oxide overlap capacitances, and the depletion layer capacitances for the S/D p-n junctions as described in Section 4.7. Thus, the gate-to-source capacitance is the sum of the

$$C_{gs} = \kappa_s\left(V_{GS}, V_{DS}, V_{BS}\right)\frac{width \cdot length \cdot \varepsilon_{ox}}{TOX} + width \cdot CGSO, \tag{4.58}$$

where $\kappa_s(V_{GS}, V_{DS}, V_{BS})$ is a scale factor between 0 and 1, which accounts for the dependence of the gate-to-source oxide capacitance on the mode of operation and bias voltages, and CGSO is the gate-to-source overlap capacitance per unit gate width. The gate-to-drain and gate-to-body capacitances are calculated in similar manner using

$$C_{gd} = \kappa_d(V_{GS}, V_{DS}, V_{BS}) \frac{width \cdot length \cdot \varepsilon_{ox}}{TOX} + width \cdot CGDO\,, \tag{4.59}$$

and

$$C_{gb} = \kappa_b(V_{GS}, V_{DS}, V_{BS}) \frac{width \cdot length \cdot \varepsilon_{ox}}{TOX} + width \cdot CGBO\,, \tag{4.60}$$

but with distinct voltage-dependent scale factors.

The S/D junction capacitances are also voltage-bias dependent and are calculated using

$$C_{bd} = \frac{CJ \cdot AD}{(1 - V_{BD} / PB)^{MJ}} + \frac{CJSW \cdot PD}{(1 - V_{BD} / PB)^{MJSW}} \tag{4.61}$$

and

$$C_{bs} = \frac{CJ \cdot AS}{(1 - V_{BS} / PB)^{MJ}} + \frac{CJSW \cdot PS}{(1 - V_{BS} / PB)^{MJSW}}\,, \tag{4.62}$$

where AS and AD are the areas of the source and drain, and PS and PD are the perimeters of the source and drain. AS, AD, PS, and PD are specified for each particular device rather than in the model statement.

4.9.2 Berkeley Short-Channel Insulated Gate Field Effect Transistor Model

The BSIM, available in several versions, accounts for subthreshold conduction, field-dependent mobility, mobility reduction attributable to the vertical field, velocity saturation, SCEs and NCEs, channel length modulation, bias dependence of the depletion layer charge under the gate, threshold voltage roll-off, non-uniform doping effects, and DIBL. The following subsections will give a brief introduction to the BSIM1 model.

There are now four generations of BSIM models as well as a BSIMSOI model for SOI transistors. The BSIM4 MOSFET model incorporates a number of refinements that make it applicable to sub-100 nm transistors and radio frequency analog circuits. For complete descriptions of these models, the reader is referred to the BSIM user manuals provided by the University of California, Berkeley [17].

4.9.2.1 BSIM1 Parameters

Many of the BSIM1 parameters are adjusted for the effective channel length and width. In general, an electrical parameter Z' is determined using

$$Z' = Z + \frac{LZ}{L - DL} + \frac{WZ}{W - DW}, \tag{4.63}$$

where the effective channel length is $L_{eff} = L - DL$, and the effective channel width is $W_{eff} = W - DW$. L and W are the *drawn* channel length and channel width, as determined the lithography, whereas DL and DW are the reductions of the length and width in the physical device attributable to lateral doping effects and encroachment. This approach has been used because DL and DW generally do not scale with L and W.

Table 4.10 summarizes the BSIM1 parameters and their coefficients for L, W variation. Not all of the BSIM1 parameters bear a clear connection to the underlying device physics; in fact, for short-channel devices, the BSIM1 parameters may lose all physical significance to become merely fitting parameters based on measured electrical characteristics.

TABLE 4.10

BSIM1 Model Parameters

BSIM1 parameter	Description	Units	L , W variation	
VFB	Flat band voltage	V	LVFB	WVFB
PHI	Surface inversion potential	V	LPHI	WPHI
K1	Body effect coefficient	$V^{1/2}$	LK1	WK1
K2	Darin/source depletion charge sharing coefficient		LK2	WK2
ETA	Zero-bias drain-induced barrier lowering coefficient		LETA	WETA
MUZ	Zero-bias mobility	cm^2/Vs		
DL	Shortening of channel	μm		
DW	Narrowing of channel	μm		
UO	Zero-bias transverse field mobility degradation coefficient	V^{-1}	LUO	WUO
U1	Zero-bias velocity saturation coefficient	μm/V	LU1	WU1
X2MZ	Sensitivity of mobility to substrate bias at $V_{DS} = V_{DD}$	cm^2/Vs	LX2MZ	WX2MZ

(Continued)

TABLE 4.10 (Continued)

BSIM1 parameter	Description	Units	L , W variation	
X2E	Sensitivity of DIBL to substrate bias	V^{-1}	LX2E	WX2E
X3E	Sensitivity of DIBL to drain bias at $V_{DS} = V_{DD}$	V^{-1}	LX3E	WX3E
X2U0	Sensitivity of transverse field mobility degradation to effective substrate bias	V^{-2}	LX2U0	WX2U0
X2U1	Sensitivity of velocity saturation effect to substrate bias	μm/V^2	LX2U1	WX2U1
MUS	Mobility at zero substrate bias with $V_{DS} = V_{DD}$	cm^2/Vs	LMUS	WMUS
X2MS	Sensitivity of mobility to substrate bias at $V_{DS=}V_{DD}$	cm^2/Vs	LX2MS	WX2MS
X3MS	Sensitivity of mobility to drain bias at $V_{DS} = V_{DD}$	cm^2/Vs	LX3MS	WX3MS
X3U1	Sensitivity of velocity saturation effect on drain bias at $V_{DS} = V_{DD}$	μm/V^2	LX3U1	WX3U1
TOX	Gate oxide thickness	μm		
TEMP	Temperature at which parameters were measured	^{o}C		
VDD	Supply voltage for measurements	V		
CGSO	Gate-source overlap capacitance per unit channel width	F/m		
CGBO	Gate-body overlap capacitance per unit channel width	F/m		
XPART*	Gate-oxide capacitance charge sharing flag			
N0	Zero-bias subthreshold slope coefficient		LN0	WN0
NB	Sensitivity of subthreshold slope to body bias		LNB	WNB
ND	Sensitivity of subthreshold slope to drain bias		LND	WND
RSH	S/D sheet resistance	Ω/square		
JS	Reverse saturation current density for S/D junctions	A/m^2		
PB	Built-in potential for S/D junctions	V		
MJ	Grading coefficient for S/D junctions			
PBSW	Built-in potential for S/D sidewall junctions	V		
MJSW	Grading coefficient for S/D sidewall junctions			
CJ	Zero-bias S/D junction capacitance per unit area	F/m^2		
CJSW	Zero bias S/D sidewall junction capacitance per unit length	F/m		
WDF	S/D default width	m		
DELL	S/D junction length reduction	m		

* If XPART = 0, then the S/D saturation charge partitioning is 60%/40%. If XPART = 1, then the partitioning is 100%/0%.

4.9.2.2 BSIM1 Threshold Voltage

The threshold voltage is calculated by

$$V_{TH} = VFB' + PHI' + K1' \cdot \sqrt{PHI' + V_{SB}} - K2' \cdot \left(PHI' + V_{SB}\right) - ETADB' \cdot V_{DS}, \tag{4.64}$$

where

$$ETADB' = ETA' - X2E' \cdot V_{SB} + X3E' \cdot \left(V_{DS} - VDD\right). \tag{4.65}$$

Accurate determination of the threshold voltage is very important because of its strong effect on the drain current, especially in low-voltage circuits.

4.9.2.3 BSIM1 Drain Current-Linear Region

In the linear region of operation, the drain current is calculated by

$$I_D = \frac{MU0'}{\left[1 + U0Z' \cdot \left(V_{GS} - V_{TH}\right)\right]} \cdot \frac{C_{ox}\left(W - DW\right)/\left(L - DL\right)}{1 + V_{DS} \cdot U1Z'/\left(L - DL\right)} \cdot \left[\left(V_{GS} - V_{TH}\right)V_{DS} - aV_{DS}^2/2\right], \tag{4.66}$$

where

$$a = 1 + \frac{gK1'}{2\sqrt{PHI' + V_{SB}}} \tag{4.67}$$

and

$$g = 1 - \frac{1}{1.744 + 0.8364\left(PHI' + V_{SB}\right)}. \tag{4.68}$$

The mobility parameter $MU0'$ is found by interpolation using the values of $MU0$ for $V_{DS} = 0$ and $V_{DS} = V_{DD}$:

$$MUO\left(V_{DS} = 0\right) = MUZ' - X2MZ' \cdot V_{SB} \tag{4.69}$$

and

$$MUO\left(V_{DS} = V_{DD}\right) = MUS' - X2MS' \cdot V_{SB}. \tag{4.70}$$

The mobility degradation parameters are found by

$$U0Z' = U0' - X2U0' \cdot V_{SB} \tag{4.71}$$

and

$$U1Z' = U1' - X2U1' \cdot V_{SB} + U3U1' \cdot \left(V_{DS} - VDD\right). \tag{4.72}$$

4.9.2.4 BSIM1 Drain Current-Saturation Region

In the linear region of operation, the drain current is calculated by

$$I_D = \frac{MU0'}{\left[1+U0Z'\cdot\left(V_{GS}-V_{TH}\right)\right]}\cdot\frac{C_{ox}\left(W-DW\right)/\left(L-DL\right)}{2aK}\cdot\left(V_{GS}-V_{TH}\right)^2, \quad (4.73)$$

where

$$K = \frac{1+v_c+\sqrt{1+2v_c}}{2}, \quad (4.74)$$

$$v_c = \frac{U1Z'}{L-DL}\cdot\frac{\left(V_{GS}-V_{TH}\right)}{a}, \quad (4.75)$$

and

$$a = 1+\frac{gK1'}{2\sqrt{PHI'+V_{SB}}} \quad (4.76)$$

4.9.2.5 BSIM1 Drain Current-Subthreshold Region

In the subthreshold, or *weak inversion*, region for which $V_{GS} < V_{TH}$, the drain current is calculated using

$$I_D = \frac{I_{\text{exp}}\cdot I_{\text{limit}}}{I_{\text{exp}}+I_{\text{limit}}}, \quad (4.77)$$

where

$$I_{\text{exp}} = MU0'\cdot C_{ox}\cdot\frac{W-DW}{L-DL}\left(\frac{kT}{q}\right)^2 \exp(1.8)\exp\left(\frac{q\left(V_{GS}-V_{TH}\right)}{N'kT}\right)\left[1-\exp\left(\frac{-qV_{DS}}{kT}\right)\right], \quad (4.78)$$

$$I_{\text{limit}} = \frac{MU0'\cdot C_{ox}}{2}\cdot\frac{W-DW}{L-DL}\left(\frac{3kT}{q}\right)^2, \quad (4.79)$$

and the subthreshold slope parameter is given by

$$N' = N0'-NB'\cdot V_{SB}+ND'\cdot V_{DS}. \quad (4.80)$$

4.9.2.6 Hand Calculations Related to the BSIM1

The complexity of the BSIM drain current equations precludes their use for hand calculations and also obscures the relationships between device design and circuit performance. However, good design is nearly always based on a clear understanding of these relationships and their use as a guide when making design improvements. Otherwise, the designer must rely on a time-consuming trial and error approach that tends to yield an inferior solution. Here simple approximations are given to facilitate hand analysis and design using short-channel devices with a BSIM1 description.

The threshold voltage may be estimated from

$$V_T \approx VFB + PHI + K1 \cdot \sqrt{PHI + V_{SB}} - K2 \cdot (PHI + V_{SB}) \,, \tag{4.81}$$

and the drain current may be estimated using the set of equations

$$I_D \approx \frac{MUZ \cdot C_{ox} \cdot W}{L} \cdot \left[(V_{GS} - V_T)V_{DS} - V_{DS}^2 / 2\right] \cdot \left[1 + (U1Z / L)V_{DS}\right] \text{(linear)}, \tag{4.82}$$

$$I_D \approx \frac{MUZ \cdot C_{ox} \cdot W}{2L} \cdot (V_{GS} - V_T)^2 \cdot \left[1 + (U1Z / L)V_{DS}\right] \text{ (saturation)}, \tag{4.83}$$

and

$$I_D \approx \frac{MUZ \cdot C_{ox} \cdot W}{L} \left(\frac{kT}{q}\right)^2 \exp(1.8) \cdot \exp\left(\frac{q(V_{GS} - V_T)}{N0 \cdot kT}\right) \text{ (subthreshold)}. \tag{4.84}$$

By relating these equations to the simple hand analysis equations developed in the previous sections of this chapter, we find that the approximate process transconductance parameter is

$$k' \approx MUZ \cdot C_{ox} \,, \tag{4.85}$$

and the approximate channel length modulation parameter is

$$\lambda \approx U1Z / L \,. \tag{4.86}$$

The approximate subthreshold parameter is

$$m \approx N0 \,. \tag{4.87}$$

4.10 SPICE Demonstrations

For the purpose of illustration, simulations were performed using Cadence Capture CIS 10.1.0 PSpice (Cadence Design Systems, San Jose, CA). The level 1 MOS transistor model parameters given in Tables 4.11 and 4.12 were used unless otherwise noted. The process transconductance parameters were calculated assuming an oxide thickness of 9 nm. For n-MOSFETS,

$$KP = \frac{(3.9)\left(8.85\times 10^{-14} F/cm\right)\left(580 cm^2 V^{-1} s^{-1}\right)}{9\times 10^{-7} cm} = 222\mu A/V^2, \tag{4.88}$$

and for p-MOSFETS,

$$KP = \frac{(3.9)\left(8.85\times 10^{-14} F/cm\right)\left(230 cm^2 V^{-1} s^{-1}\right)}{9\times 10^{-7} cm} = 88\mu A/V^2. \tag{4.89}$$

The overlap capacitances per unit gate width were determined with the assumption that $L_{OV} = 0.1\mu m$:

$$\begin{aligned} CGSO &= \frac{(3.9)\left(8.85\times 10^{-14} F/cm\right)\left(0.1\times 10^{-4} cm\right)}{9\times 10^{-7} cm} \\ &= 3.8 pF/cm = 0.38 nF/m \end{aligned} \tag{4.90}$$

and

$$\begin{aligned} CGDO &= \frac{(3.9)\left(8.85\times 10^{-14} F/cm\right)\left(0.1\times 10^{-4} cm\right)}{9\times 10^{-7} cm}. \\ &= 3.8 pF/cm = 0.38 nF/m \end{aligned} \tag{4.91}$$

TABLE 4.11

n-MOS Level 1 SPICE Parameters

Parameter	Value	Units
KP	222u	A/V^2
VTO	0.5	V
GAMMA	0.15	V$^{1/2}$
PHI	0.7	V
LAMBDA	0.05	
TOX	9n	m
NSUB	1E16	cm^{-3}
UO	580	cm^2/Vs
CGSO	0.38n	F/m
CGDO	0.38n	F/m

TABLE 4.12

p-MOS Level 1 SPICE Parameters

Parameter	Value	Units
KP	88u	A/V^2
VTO	−0.5	V
GAMMA	0.15	V$^{1/2}$
PHI	0.7	V
LAMBDA	0.05	
TOX	9n	m
NSUB	1E16	cm^{-3}
UO	230	cm^2/Vs
CGSO	0.38n	F/m
CGDO	0.38n	F/m

The body effect coefficient was calculated from

$$GAMMA = \frac{\sqrt{2q\varepsilon_{Si}N_a}}{C_{ox}}$$

$$= \frac{\sqrt{2\left(1.602\times10^{-19}C\right)(11.9)\left(8.85\times10^{-14}F/cm\right)\left(10^{16}cm^{-3}\right)}}{(3.9)\left(8.85\times10^{-14}F/cm\right)/9\times10^{-7}cm}$$

$$\approx 0.15V^{1/2}.$$

SPICE Example 4.1 n-MOSFET Characteristics with λ = 0

Characteristic curves (I_D as a function of V_{DS} with V_{GS} as a parameter) were developed using a DC sweep of V_{DS} and a nested parametric sweep of V_{GS} with the circuit of Figure 4.34. The channel length modulation parameter λ was set to zero for these simulations. The resulting characteristic curves in Figure 4.35 are flat within the saturation region.

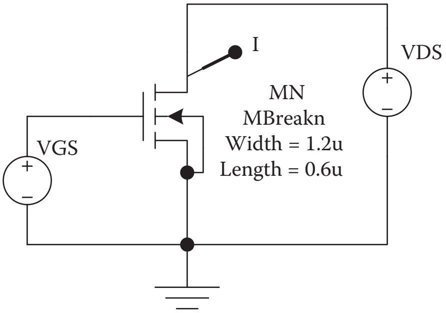


FIGURE 4.34
SPICE circuit used for the determination of the characteristic curves for an n-MOS transistor.

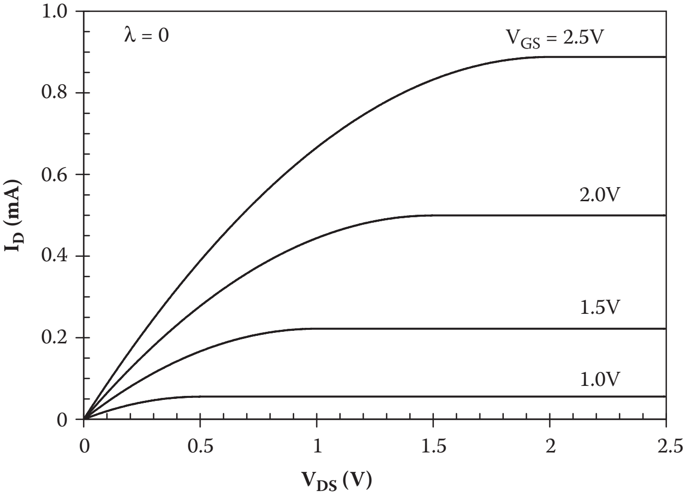


FIGURE 4.35.
Characteristic curves for the n-MOS transistor of Figure 4.34.

SPICE Example 4.2 Channel Length Modulation in an n-MOS Transistor

Figure 4.36 shows characteristic curves for a n-MOS transistor with $\lambda = 0.05$. Accounting for the channel length, modulation increases the drain current by at most 12.5% here, and usually we will neglect the channel length modulation for the purpose of approximate hand calculations.

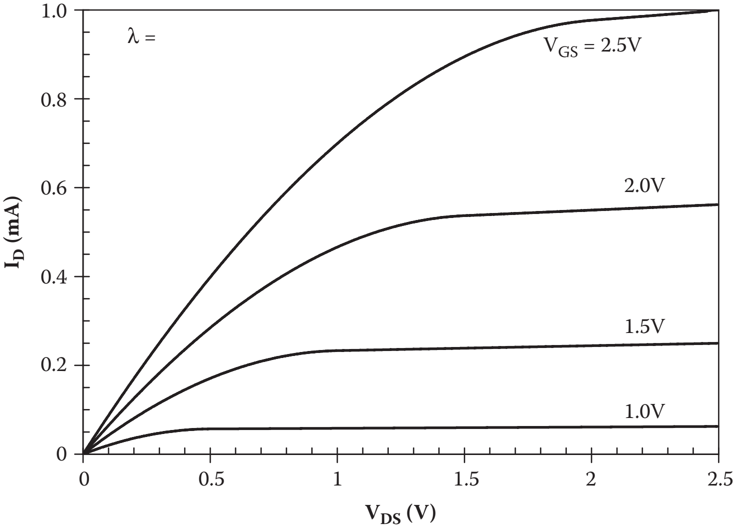


FIGURE 4.36
n-MOS transistor characteristics for the case of $\lambda = 0.05$.

SPICE Example 4.3 Body Effect in an n-MOS Transistor

To explore the body effect in an n-MOS transistor, the circuit of Figure 4.37 was used with a DC sweep of the gate-to-source voltage and a parametric sweep of the body-to-source bias. As can be seen from Figure 4.38, the application of a negative bias on the body makes the threshold voltage more positive. Also, for saturated operation at a given value of V_{GS}, a negative body-to-source voltage decreases the drain current.

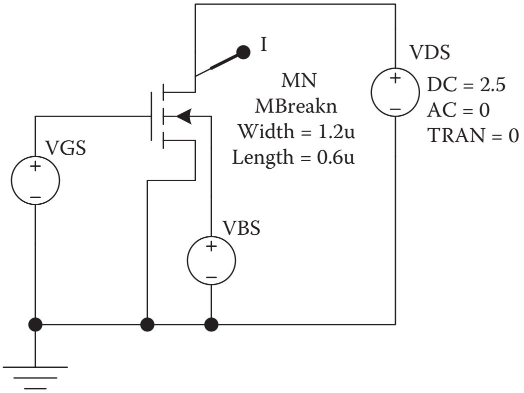


FIGURE 4.37
Circuit for the determination of the drain current as a function of V_{GS} with V_{BS} as a parameter.

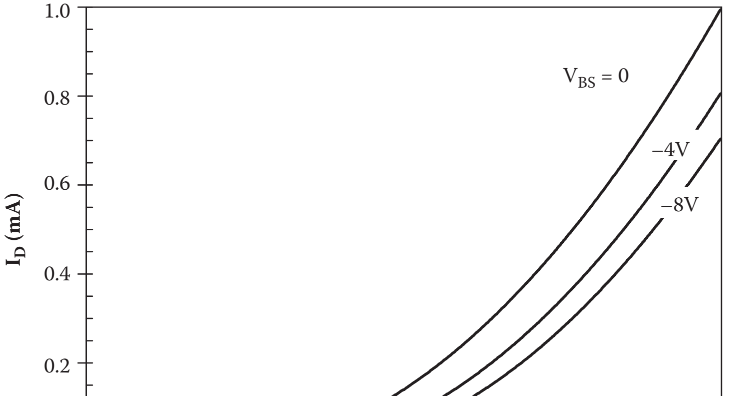


FIGURE 4.38
Drain current as a function of the drain-to-source voltage with body-to-source bias as a parameter for an n-MOS transistor.

SPICE Example 4.4 p-MOSFET Characteristics

Characteristic curves were determined for a p-MOS transistor using the circuit of Figure 4.39 with a DC sweep of V_{DS} and a parametric sweep of V_{GS}. All values of V_{GS} and V_{DS} are negative, but the drain current flowing out of the drain is considered positive. As shown in Figure 4.40, the maximum saturated drain current (~0.4 mA) is less than in the case of the n-MOS transistor (~1.0 mA) because of the lower process transconductance parameter.

4.11 Practical Perspective

For practical perspective articles, see the dynamic website at http://www.engr.uconn.edu/ece/books/ayers.

4.12 Summary

The MOSFET (or MOS transistor) is presently the most important device for digital integrated circuits. CMOS integrated circuits use complementary pairs of enhancement-type (normally off) n-MOS and p-MOS transistors. The threshold voltages for these devices may be determined based on the metal-semiconductor work function difference, the doping in the semiconductor, and the charge in the gate insulator.

An MOS transistor can operate in one of three modes: linear, saturation, and cutoff. The current versus voltage characteristics for a long-channel MOS transistor may be calculated based on the assumption that carriers

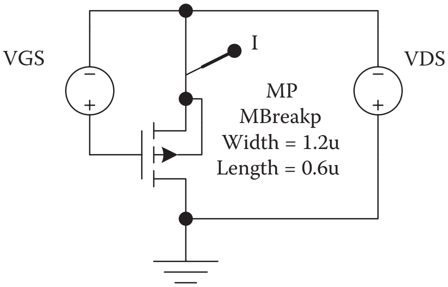


FIGURE 4.39
Circuit used for the determination of the characteristics for a p-MOS transistor.

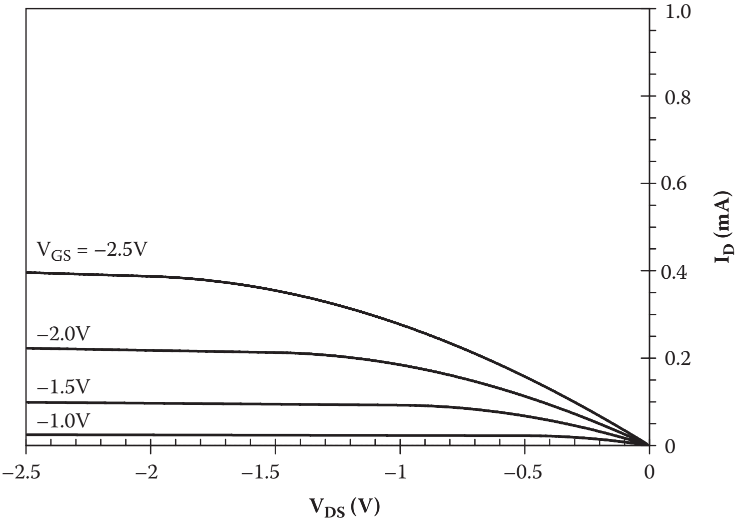


FIGURE 4.40
Characteristics for a p-MOS transistor.

drift according to their low-field mobilities. For short-channel MOSFETs, it is necessary to account for carrier velocity saturation. In the subthreshold (cutoff) region of operation, current flow is attributable to diffusion of minority carriers from the source to the drain. This mode of operation is important in determining the standby dissipation in VLSI circuits.

The capacitances in an MOS transistor include contributions attributable to the gate oxide and also the p-n junctions formed at the source and drain. These capacitances are important in determining the delay times of MOS circuits.

Short-channel MOS transistors behave differently to long-channel devices in several important ways. As a consequence of the SCE, the threshold voltage is decreased in absolute value compared with a similar long-channel device. There is also an NCE, which changes the threshold voltage in the opposite sense and compensates partly for the SCE. The current-voltage characteristics are influenced strongly by carrier velocity saturation, as mentioned above. DIBL causes a degradation of the subthreshold characteristics and an increase in the subthreshold current.

SPICE modeling is necessary for accurate performance predictions of MOS digital circuits. The SPICE level 1 MOSFET model is closely related to the simple electrical models used for hand analysis of MOS circuits. The BSIM is considerably more complex but provides improved accuracy for short-channel MOS transistors; it is therefore used extensively in industry.

4.13 Exercises

E4.1. An n-MOS transistor is fabricated with $t_{ox} = 6\text{nm}$ and $N_a = 1.2 \times 10^{16} cm^{-3}$. The gate is heavily doped n-polysilicon (with the Fermi level coincident with the conduction band), and there are $9 \times 10^{10} cm^{-2}$ positive charges in the oxide. What is the zero-bias threshold voltage for this device? Is it an enhancement-type or depletion-type device?

E4.2. Find the zero-bias threshold voltage for an n-MOS transistor with $t_{ox} = 8\text{nm}$ and $N_a = 9 \times 10^{15} cm^{-3}$. The gate is heavily doped n-polysilicon (with the Fermi level coincident with the conduction band), there are $1.1 \times 10^{11} cm^{-2}$ positive charges in the oxide, and a boron dose of $1.2 \times 10^{12} cm^{-2}$ is implanted to adjust the threshold voltage.

E4.3. An n-MOS transistor is fabricated with $t_{ox} = 5\text{nm}$ and $N_a = 1.3 \times 10^{16} cm^{-3}$. The gate is heavily doped n-polysilicon (with the Fermi level coincident with the conduction band), and there are $1.0 \times 10^{11} cm^{-2}$ positive charges in the oxide. Describe the necessary ion implantation step (type of impurity and dose) necessary to produce a depletion-type device with a threshold voltage of –0.3 V.

E4.4. A metal gate n-MOS transistor has $\phi_{ms} = -0.7V$. If $t_{ox} = 8\text{nm}$, $N_a = 12 \times 10^{16}$ cm^{-3}, and there are $0.9 \times 10^{11} cm^{-2}$ positive charges in the oxide, what is the necessary boron ion implantation dose to shift the zero-bias threshold voltage to 0.4 V?

E4.5. Calculate the zero-bias threshold voltage for a p-channel MOSFET with $t_{ox} = 6\text{nm}$ 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 there are $10^{11} cm^{-2}$ positive charges in the oxide. A phosphorus dose of $5 \times 10^{11} cm^{-2}$ is implanted to adjust the threshold voltage.

E4.6. Consider a p-MOS transistor with $t_{ox} = 7\text{nm}$ and $N_d = 8 \times 10^{15} cm^{-3}$. Determine the required ion implantation (impurity and dose) to shift the zero-bias threshold by –0.2 V.

E4.7. Consider a p-channel MOSFET with a metal gate having $\phi_{ms} = +0.6V$, $t_{ox} = 5nm$, and $N_d = 10^{16} cm^{-3}$. There are $1.1 \times 10^{11} cm^{-2}$ positive charges in the oxide. Determine the necessary ion implantation step (type of impurity and dose) to shift the zero-bias threshold voltage to –0.3 V.

E4.8. An n-MOS transistor is fabricated with $t_{ox} = 6nm$ and $N_a = 1 \times 10^{16} cm^{-3}$. The gate is heavily doped n-polysilicon (with the Fermi level coincident with the conduction band), there are $1.0 \times 10^{11} cm^{-2}$ positive charges in the oxide, and a boron dose of $1.5 \times 10^{12} cm^{-2}$ is implanted to adjust the threshold voltage. Determine the body bias V_{BS} necessary to shift the threshold voltage to +0.5 V.

E4.9. A p-MOS transistor is fabricated with $t_{ox} = 5nm$ and $N_d = 1 \times 10^{16} cm^{-3}$. The gate is heavily doped n-polysilicon (with the Fermi level coincident with the conduction band), there are $1.0 \times 10^{11} cm^{-2}$ positive charges in the oxide, and a phosphorus dose of $1.5 \times 10^{12} cm^{-2}$ is implanted to adjust the threshold voltage. Find the threshold voltage with a body bias $V_{BS} = 2.5V$.

E4.10. Consider n-channel and p-channel silicon MOSFETs fabricated on the same wafer with channel lengths of 1.0-μm-thick 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-MOS and p-MOS transistors such that the device transconductance parameters are both 0.5 mA/V^2

E4.11. An n-MOS transistor with dimensions $L_N = 0.6\mu m$ and $W_N = 2.4\mu m$. If the device transconductance parameter is determined to be 0.9 mA/V^2, what is the approximate oxide thickness?

E4.12. For an n-MOS transistor with $t_{ox} = 5nm$, $L_N = 0.1\mu m$, and $W_N = 0.2\mu m$, calculate the saturated drain current with $V_{GS} - V_{TN} = 0.5V$ based on (1) the long-channel equation and (2) the short-channel equation. Which is more appropriate?

E4.13. For a p-MOS transistor with $t_{ox} = 5nm$, $L_P = 0.1\mu m$, and $W_P = 0.2\mu m$, calculate the saturated drain current with $V_{GS} - V_{TP} = -0.5V$ based on (1) the long-channel equation and (2) the short-channel equation. Which is more appropriate?

E4.14. Calculate the characteristic curves for an n-MOS transistor $t_{ox} = 5nm$, $L_N = 0.6\mu m$, $W_N = 1.2$ μm, and $V_{TN} = 0.3V$ assuming that the long-channel equations are applicable.

E4.15. Calculate the characteristic curves for an n-MOS transistor $t_{ox} = 5nm$, $L_N = 0.1\mu m$, $W_N = 0.3\mu m$, and $V_{TN} = 0.3V$ using the short-channel equations.

E4.16. Find the subthreshold power dissipation in an n-MOS transistor with $t_{ox} = 6nm$, $L_N = 0.25\mu m$, $W_N = 0.5\mu m$, $V_{TN} = 0.3V$, $V_{DS} = 1.5V$, and $V_{GS} = 0V$. The subthreshold swing is 95 mV.

E4.17. Plot the subthreshold current as a function of the gate-to-source bias for an n-MOS transistor with $t_{ox} = 4nm$, $L_N = 0.15\mu m$,, $W_N = 0.3\mu m$, $V_{TN} = 0.3V$, and $V_{DS} >> 3kT/q$, if the subthreshold swing is 100 mV.

E4.18. Estimate the transit time for an n-channel MOSFET with $L = 45nm$ and V_{DS} = 1V, assuming (1) the constant mobility model and (2) the velocity saturation model. Which model is more appropriate?

E4.19. Create the layout design for an n-channel MOSFET so that $I_{Dsat} = 2mA$ with $V_{GS} = 2.5V$. $t_{ox} = 10nm$, $V_{TN} = 0.5V$, and $2X = 1\mu m \cdot$

E4.20. Create the layout design for a p-MOS transistor so that $I_{Dsat} = 2mA$ with $V_{GS} = -2.5V$. $t_{ox} = 10nm$, $V_{TP} = -0.5V$, and $2X = 1\mu m \cdot$

E4.21. Create the layout design for an n-channel MOSFET so that $I_{Dsat} = 2mA$ with $V_{GS} = 1.0V$. $t_{ox} = 4nm$, $V_{TN} = 0.3V$, and $2X = 0.1\mu m$. Set the gate length equal to 2X and use the short-channel equation for I_{Dsat}.

E4.22. Create the layout design for a p-MOSFET so that $I_{Dsat} = 2mA$ with $V_{GS} = 1.0V$. $t_{ox} = 4nm$, $V_{TP} = -0.3V$, and $2X = 0.1\mu m$. Set the gate length equal to 2X and use the short-channel equation for I_{Dsat}.

E4.23. An n-MOS transistor has dimensions $t_{ox} = 4nm$, $W = 0.5\mu m$, $L = 0.25\mu m$, $L_{OV} = 0.05\mu m$, $L_D = L_S = 1.0\mu m$, and $x_j = 0.05\mu m$. The substrate doping concentration is $N_a = 10^{16} cm^{-3}$, the sidewall (channel stopper) doping is $N_a = 5 \times 10^{16} cm^{-3}$, and the S/D regions are doped to the concentration of $N_d = 10^{18} cm^{-3}$. Find the worst-case capacitance between the gate and ground.

E4.24. An n-MOS transistor has dimensions $t_{ox} = 4nm$, $W = 0.5\mu m$, $L = 0.25\mu m$, $L_{OV} = 0.05\mu m$, $L_D = L_S = 1.0\mu m$, and $x_j = 0.05\mu m$. The substrate doping concentration is $N_a = 10^{16} cm^{-3}$, the sidewall (channel stopper) doping is $N_a = 5 \times 10^{16} cm^{-3}$, and the S/D regions are doped to the concentration of $N_d = 10^{18} cm^{-3}$. Estimate the zero-bias capacitance between the drain and ground.

For additional exercise problems, see the dynamic website at http://www.engr.uconn.edu/ece/books/ayers.

References

1. Nicollian, E.H., and Brews, J.R., *MOS physics and technology*, Wiley, New York, 1982.
2. Brews, J.R., *Physics of the MOS transistor,* Applied Solid State Science, Supplement 2A, edited by Kahng, D., Academic, New York, 1981.
3. Taur, Y., and Ning, T.H., *Fundamentals of modern VLSI devices*, Cambridge University Press, New York, 1998.
4. Pao, H.C., and Sah, C.T., Effects of diffusion current on characteristics of metal-oxide (insulator)-semiconductor transistors. *Solid-State Electron.*, 9, 927, 1966.
5. Brews, J.R., A charge sheet model of the MOSFET. *Solid-State Electron.*, 21, 345, 1978.
6. Yau, L.D., A simple theory to predict the threshold voltage of short-channel IGFETs. *Solid-State Electron.*, 17, 1059, 1974.
7. Troutman, R.R., VLSI limitations from drain-induced barrier lowering. *IEEE Trans. Electron. Dev.*, ED-26, 461, 1979.
8. 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.
9. Caughey, D.M., and Thomas, R.E., Carrier mobilities in silicon empirically related to doping and field. *Proc. IEEE*, 55, 2192, 1967.
10. Taylor, G.W., Velocity saturated characteristics of short-channel MOSFETs. *Bell Labs. Tech. J.*, 63, 1325, 1984.

11. 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 Dev.*, 49, 832, 2002.
12. Hu, C., BSIM model for circuit design using advanced technologies, Digest of Technical Papers 2001, *Symp. VLSI Circuits*, 5, 2001.
13. 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 Dev.*, 277, 1997.
14. 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.
15. Arora, N., *MOSFET models for VLSI circuit simulation*, Springer-Verlag, Vienna, 1993.
16. 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.
17. University of California, Berkeley, CA, http://www.berkeley.edu.

5

MOS Gate Circuits

5.1 Inverter Static Characteristics

The simplest MOS gate circuit is an inverter made using a single n-channel MOSFET switch with a pull-up device as shown in Figure 5.1. The pull-up device could be a resistor, as in Figure 5.1a, or an active device, as in Figure 5.1c.

The output voltage versus input voltage characteristic (voltage transfer characteristic) may be determined by equating the current in the switch MOSFET with the current in the pull-up device. For example, in the case of a MOS inverter with a resistor pull-up device, the drain current I_{DO} in the MOSFET switch is given by

$$I_{DO} = \begin{cases} 0; \\ \mathrm{K_O}\left[(V_{IN} - V_{TO})V_{OUT} - V_{OUT}^2/2\right]; \\ K_{NO}(V_{IN} - V_{TN})^2/2; \end{cases}$$

$$\begin{array}{rl} 0 \le V_{IN} \le V_T & \text{(cutoff)} \\ V_T \le V_{IN} \le (V_{OUT} + V_T) & \text{(linear)} \\ (V_{OUT} + V_T) \le V_{IN} & \text{(saturation)} \end{array} \tag{5.1}$$

where K_O and V_{TO} are the device transconductance parameter and threshold voltage for the switch device $\mathrm{M_{NO}}$, and the resistor current is given by

$$I_L = \frac{V_{DD} - V_{OUT}}{R}, \tag{5.2}$$

where V_{DD} is the supply voltage, and R is the value of the load resistor. The voltage transfer characteristic, $\mathrm{V_{OUT}}$ as a function of $\mathrm{V_{IN}}$, may be determined from

$$I_{DO}(V_{IN}, V_{OUT}) = I_L(V_{IN}, V_{OUT}). \tag{5.3}$$

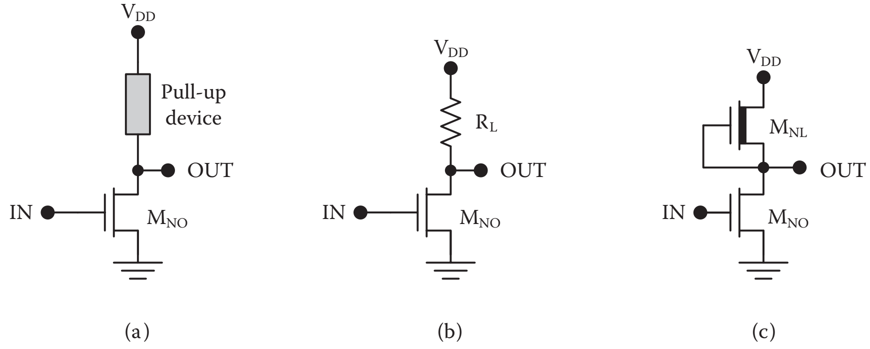


FIGURE 5.1
MOS inverter designs: (a) General MOS inverter using an n-channel MOSFET switch and a pull-up device, (b) MOS inverter using a pull-up resistor, and (c) MOS inverter using a depletion mode n-channel MOSFET as the pull-up device.

Analytical solutions can be found for each of the three regimes given in Equation 5.1.

It is also possible to determine the drain current and output voltage by a graphical method referred to as the "load curve analysis." Here, the drain current I_{DO} and the pull-up resistor current I_L are both plotted as functions of V_{OUT}, for a particular value of V_{IN}. The intersection of the MOSFET characteristic with the "load curve" provides the solution. Figure 5.2 shows an example of this load curve analysis applied to an MOS inverter with a resistive load, for the case of $V_{IN} = 1.0V$. The circuit parameters are $V_{DD} = 2.5V$, $K_{NO} = 100\mu A / V^2$, $V_{TO} = 0.5V$, and $R = 50k\Omega$. The solution, found from the intersection of the transistor characteristic and the load curve, is $I_{DD} = 25\mu A$ and $V_{DD} = 1.25V$.

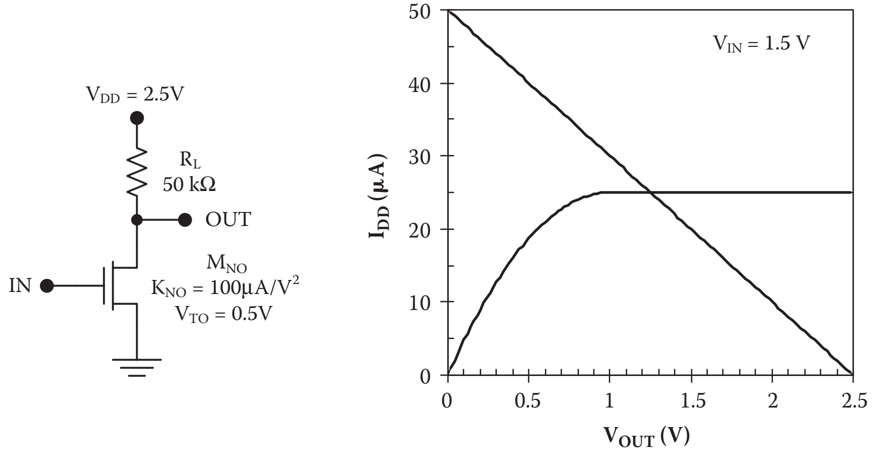


FIGURE 5.2
Load curve analysis for a resistor-loaded MOS inverter.

The graphical analysis may be extended to a range of input voltages to determine the voltage transfer characteristic (V_{OUT} as a function of V_{IN}); this *load surface* analysis is illustrated in Figure 5.3. Here, the drain current is plotted as a function of both V_{IN} and V_{OUT} for both the pull-down and pull-up devices. The intersection of these two surfaces provides the solution, and its projection onto the voltage plane is the voltage transfer characteristic (shown in Figure 5.4).

This same graphical analysis can be applied to the MOS inverter with a depletion-type load as shown in Figure 5.1c. As in the previous circuit, the drain current I_{DO} in the MOSFET switch is given by

$$I_{DO} = \begin{cases} 0; \\ \mathrm{K_O}\left[(V_{IN} - V_{TO})V_{OUT} - V_{OUT}^2/2\right]; \\ K_O(V_{IN} - V_{TO})^2/2; \end{cases}$$

$$\begin{array}{rl} 0 \le V_{IN} \le V_T & \text{(cutoff)} \\ V_T \le V_{IN} \le (V_{OUT} + V_T) & \text{(linear)} \\ (V_{OUT} + V_T) \le V_{IN} & \text{(saturation)} \end{array} \qquad (5.4)$$

where K_O and V_{TO} are the device transconductance parameter and threshold voltage for the switch device M_{NO}. The drain current in the pull-up device is given by

$$I_{DL} = \begin{cases} \mathrm{K_L} V_{TL}^2/2; \\ \mathrm{K_L}\left[(-V_{TL})(V_{DD} - V_{OUT}) - (V_{DD} - V_{OUT})^2/2\right]; \end{cases}$$

$$\begin{array}{rl} V_{OUT} < (V_{DD} + V_{TL}) & \text{(saturation)} \\ V_{OUT} > (V_{DD} + V_{TL}) & \text{(linear)} \end{array} \qquad (5.5)$$

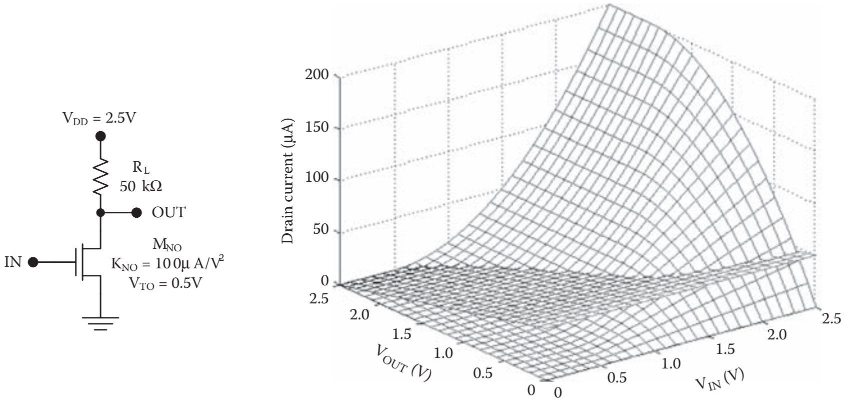


FIGURE 5.3
Load surface analysis for a resistor-loaded MOS inverter.

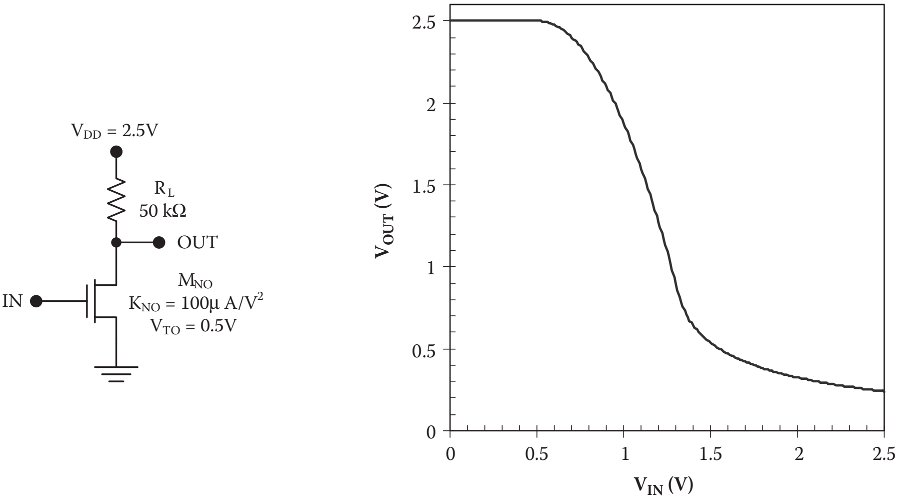


FIGURE 5.4
Voltage transfer characteristic for a resistor-loaded MOS inverter.

where K_L and V_{TL} are the device transconductance parameter and threshold voltage for the pull-up device M_{NL}. Figure 5.5 shows the load curve analysis with V_{IN} = 1.5 V, and Figure 5.6 shows the load surface analysis. The voltage transfer characteristic, determined from the projection of the load surface solution into the voltage plane, is shown in Figure 5.7. In this figure, the three regions are labeled according to the mode of operation for the pull-down device.

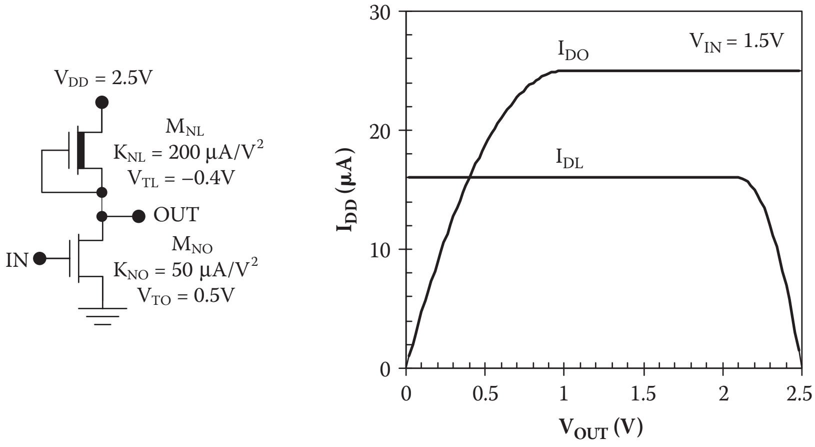


FIGURE 5.5
Load curve analysis for an MOS inverter with a depletion-type load.

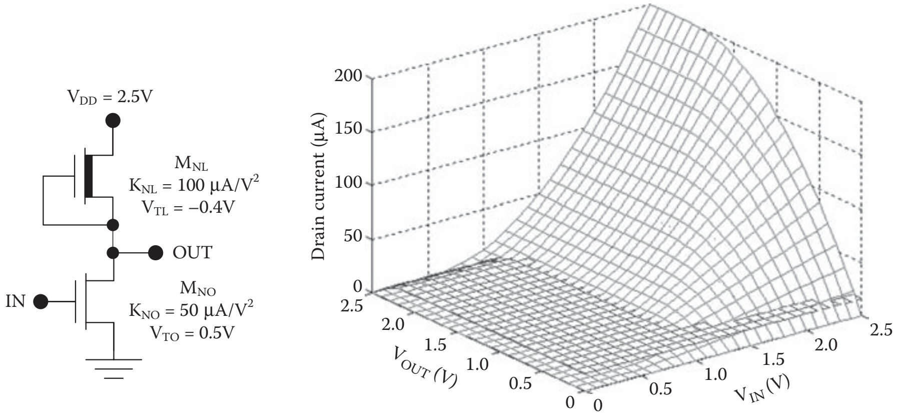


FIGURE 5.6
Load surface analysis for an MOS inverter with a depletion-type load.

5.2 Critical Voltages

Analytic expressions can be found for the critical input and output voltages of the depletion-loaded MOS inverter by equating the drain currents in the pull-down and pull-up devices. These include the output low voltage, the

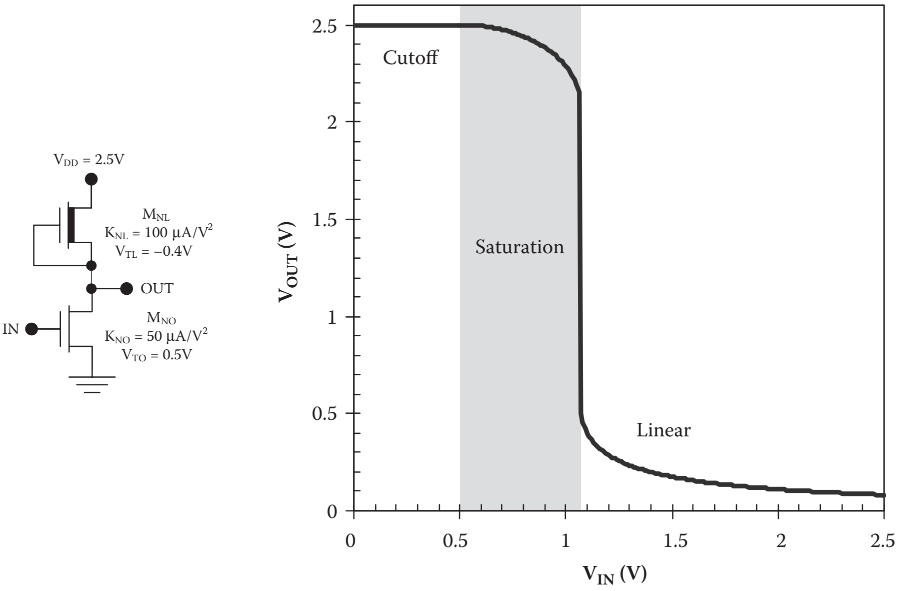


FIGURE 5.7
Voltage transfer characteristic for an MOS inverter with a depletion-type load. The mode of operation for the pull-down device is annotated in the three regions of the figure.

output high voltage, the input low voltage, the input high voltage, and the switching threshold. For this purpose, we will consider the depletion-loaded MOS inverter circuit depicted in Figure 5.8.

5.2.1 Output High-Voltage V_{OH}

With a logic zero input, the switch transistor M_{NO} is cutoff. For the load device M_{NL}, the drain current is zero, and the device is in the ohmic region of operation. Therefore, the drain to source voltage for the load is zero and

$$V_{OH} = V_{DD} \tag{5.6}$$

5.2.2 Output Low-Voltage V_{OL}

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, we have

$$K_O\left[(V_{DD} - V_{TO})V_{OL} - \frac{V_{OL}^2}{2}\right] = \frac{K_L}{2}V_{TL}^2, \tag{5.7}$$

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}, we obtain

$$V_{OL} = V_{DD} - V_{TO} \pm \sqrt{(V_{DD} - V_{TO})^2 - \left(\frac{K_L}{K_O}\right)V_{TL}^2}\ . \tag{5.8}$$

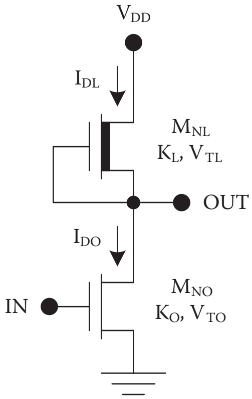


FIGURE 5.8
MOS inverter with a depletion-type pull-up device.

Notice that the quadratic formula predicts two solutions; however, the solution stemming from use of the plus sign is nonphysical and must be discarded. We can conclude that the output low voltage depends on the ratio of the device transconductance parameters but not their absolute values. Hence, we can scale both devices up or down in size without affecting the output low voltage.

5.2.3 Input Low-Voltage V_{IL}

The input low voltage for the depletion loaded MOS inverter can be determined with the assumption that M_{NO} is saturated and M_{NL} is linear. Then the drain currents can be written as

$$I_{DO} = \frac{K_O}{2}\left(V_{IN} - V_{TO}\right)^2 \tag{5.9}$$

and

$$I_{DL} = K_L\left[\left(-V_{TL}\right)\left(V_{DD} - V_{OUT}\right) - \frac{\left(V_{DD} - V_{OUT}\right)^2}{2}\right]. \tag{5.10}$$

If we equate the drain currents, then

$$I_{DO} = I_{DL}\,, \tag{5.11}$$

$$dI_{DO} = dI_{DL}\,, \tag{5.12}$$

and

$$\frac{\partial I_{DO}}{\partial V_{IN}}dV_{IN} = \frac{\partial I_{DL}}{\partial V_{OUT}}dV_{OUT}\,. \tag{5.13}$$

The slope of the transfer characteristic can therefore be determined using the partial derivatives:

$$\frac{dV_{OUT}}{dV_{IN}} = \frac{\dfrac{\partial I_{DO}}{\partial V_{IN}}}{\dfrac{\partial I_{DL}}{\partial V_{OUT}}} = \frac{K_O\left(V_{IN} - V_{TO}\right)}{K_L V_{TL} + K_L\left(V_{DD} - V_{OUT}\right)}\,. \tag{5.14}$$

By definition, this slope is –1 at the input low voltage. Therefore,

$$\left.\frac{K_O\left(V_{IN} - V_{TO}\right)}{K_L V_{TL} + K_L\left(V_{DD} - V_{OUT}\right)}\right|_{V_{IN}=V_{IL}} = -1\,. \tag{5.15}$$

Solving for V_{OUT}, we obtain

$$V_{OUT} = \frac{K_O}{K_L}\left(V_{IN} - V_{TO}\right) + V_{TL} + V_{DD}\,. \tag{5.16}$$

Substituting this result into Equation 5.15 and solving, we obtain the input low voltage:

$$V_{IL} = V_{TO} + \frac{K_L}{\sqrt{K_O K_L + K_O^2}}\left|V_{TL}\right|. \tag{5.17}$$

5.2.4 Input High-Voltage V_{IH}

For the determination of V_{IH}, we start with the assumption that M_{NO} is linear and M_{NL} is saturated. Then the drain currents are given by

$$I_{DO} = K_O\left[\left(V_{IN} - V_{TO}\right)V_{OUT} - \frac{V_{OUT}^2}{2}\right] \tag{5.18}$$

and

$$I_{DL} = \frac{K_L V_{TL}^2}{2}\,. \tag{5.19}$$

The drain currents are equal,

$$I_{DO} = I_{DL}\,, \tag{5.20}$$

so that

$$dI_{DO} = dI_{DL}\,. \tag{5.21}$$

However, I_{DL} is constant so that

$$\frac{\partial I_{DL}}{\partial V_{IN}} = \frac{\partial I_{DL}}{\partial V_{OUT}} = 0. \tag{5.22}$$

Therefore,

$$\frac{\partial I_{DO}}{\partial V_{IN}} dV_{IN} + \frac{\partial I_{DO}}{\partial V_{OUT}} dV_{OUT} = 0, \tag{5.23}$$

and the slope of the voltage transfer characteristic can be found from

$$\frac{dV_{OUT}}{dV_{IN}} = \frac{\dfrac{\partial I_{DO}}{\partial V_{IN}}}{-\dfrac{\partial I_{DO}}{\partial V_{OUT}}}\,. \tag{5.24}$$

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, we obtain

$$V_{IH} = V_{TO} + 2|V_{TL}|\sqrt{\frac{K_L}{3K_O}}\,. \tag{5.25}$$

5.2.5 Switching Threshold (Midpoint) Voltage V_M

The switching threshold voltage, also known as the midpoint voltage V_M, is the value of the input voltage for which $V_{OUT} = V_{IN}$. With this condition, the gate-to-source and drain-to-source voltages are equal for the pull-down device so it is saturated. The pull-up device is also saturated, so by equating the drain currents we have

$$\frac{K_O}{2}(V_M - V_{TO})^2 = \frac{K_L V_{TL}^2}{2}\,. \tag{5.26}$$

Solving, we obtain

$$V_M = V_{TO} - V_{TL}\sqrt{\frac{K_L}{K_O}}\,. \tag{5.27}$$

Example 5.1 Design for V_{OL}

Design a depletion load MOS inverter with $V_{DD} = 3.3V$, $V_{TO} = 0.6V$, and $V_{TL} = -0.4V$. The fabrication process uses 0.6 µm technology with t_{ox} = 9 nm and μ_n = 580 cm^2/Vs.

Solution: The transistors should be sized so that V_{OL} is several tenths of a volt less than V_{TO}, if we are to limit the "off" drain current in M_{NO} to an acceptable value. If we design for $V_{OL} \le 0.3V$, then the required ratio of device transconductance parameters is

$$\frac{K_O}{K_L} \ge \frac{V_{TL}^2}{2\left[(V_{DD} - V_{TO})V_{OL} - \dfrac{V_{OL}^2}{2}\right]} = \frac{(-0.4V)^2}{2\left[(3.3V - 0.6V)0.3V - \dfrac{(0.3V)^2}{2}\right]} \approx \frac{1}{9.5}.$$

Therefore,

$$\frac{W_O / L_O}{W_L / L_L} \geq \frac{1}{9.5},$$

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. There are other restrictions on the transistor sizing that are imposed by speed requirements as well, but to satisfy the DC constraints, we could use $K_O / K_L = 1/2$. If the channel lengths are both set to the minimum dimension, $L_O = L_L = 0.6\mu m$, then $W_O / W_L = 1/2$, and we could choose the specific values $W_O = 3\mu m$ and $W_L = 6\mu m$. Then

$$\begin{aligned} V_{OL} &= V_{DD} - V_{TO} - \sqrt{(V_{DD} - V_{TO})^2 - \left(\frac{K_L}{K_O}\right) V_{TL}^2} \\ &= 3.3V - 0.6V - \sqrt{(3.3V - 0.6V)^2 - (0.5)(-0.4V)^2} \\ &= 0.0149V. \end{aligned}$$

This meets the original design goal of $V_{OL} \leq 0.3V$.

Example 5.2 Voltage Transfer Characteristic

Determine the voltage transfer characteristic for the MOS inverter of Figure 5.9.

Solution: The process transconductance parameter for the transistors is

$$k' = \frac{\mu_n \varepsilon_{OX}}{t_{OX}} = \frac{(580cm^2 / Vs)(3.9)(8.85 \times 10^{-14} F / cm)}{9 \times 10^{-7} cm} = 0.22mA / V^2.$$

The device transconductance parameters are

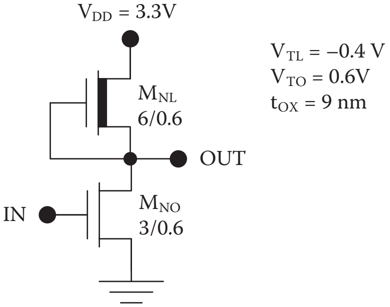


FIGURE 5.9
Example MOS inverter for the calculation of the voltage transfer characteristic.

$$K_L = k'\frac{W_L}{L_L} = \left(0.22mA/V^2\right)\left(\frac{6\mu m}{0.6\mu m}\right) = 2.2mA/V^2$$

and

$$K_O = k'\frac{W_O}{L_O} = \left(0.22mA/V^2\right)\left(\frac{3\mu m}{0.6\mu m}\right) = 1.1mA/V^2.$$

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 using the values of V_T, V_{GS}, and V_{DS} for each transistor. Specifically, M_{NL} is linear if $(V_{GSL} - V_{TL}) > V_{DSL}$ or if $V_{OUT} \geq V_{DD} + V_{TL}$; otherwise, it is in the saturation region. M_{NO} is cutoff if $V_{GSO} < V_{TO}$ or $V_{IN} < V_{TO}$. If conducting, M_{NO} is linear if $(V_{GSO} - V_{TO}) > V_{DSO}$, that is if $\left(V_{IN} + V_{TO}\right) \geq V_{OUT}$, but otherwise it is saturated.

Thus, if $V_{IN} \leq 0.6V$, M_{NO} is cutoff and M_{NL} is linear, so that $V_{OUT} = 3.3V$ $(V_{IN} \leq 0.6V)$. If $(V_{OUT} + 0.6V) \geq V_{IN} \geq 0.6V$ and $V_{OUT} \geq 2.9V$, M_{NO} is saturated and M_{NL} is linear so that

$$I_{DD} = \frac{K_O}{2}\left(V_{IN} - V_{TO}\right)^2, \text{ t}$$

and

$$V_{OUT} = V_{DD} - V_{DSL} = V_{DD} - \left[-V_{TL} - \sqrt{\left(-V_{TL}\right)^2 - \frac{K_O\left(V_{IN} - V_{TO}\right)^2}{K_L}}\right]$$

$$= 2.9V + \sqrt{\left(0.4V\right)^2 - \left(0.5\right)\left(V_{IN} - 0.6V\right)^2};$$

$$\left[\left(V_{OUT} + 0.6V\right) \geq V_{IN} \geq 0.6V\right] \text{ and } \left[V_{OUT} \geq 2.9V\right]$$

Finally, if $(V_{OUT} + 0.6\ V) \leq V_{IN}$ and $V_{OUT} \leq 2.9$ V, then M_{NO} is linear and M_{NL} is saturated so that

$$I_{DD} = \frac{K_L}{2}\left(-V_{TL}\right)^2,$$

and

$$V_{OUT} = V_{DSO} = \left(V_{IN} - V_{TO}\right) - \sqrt{\left(V_{IN} - V_{TO}\right)^2 - \frac{K_L\left(-V_{TL}\right)^2}{K_O}} -$$

$$= V_{IN} - 0.6V - \sqrt{\left(V_{IN} - 0.6V\right)^2 - \left(-0.4\right)^2/0.5};$$

$$\left[\left(V_{OUT} + 0.6V\right) \leq V_{IN}\right] \text{ and } \left[V_{OUT} \leq 2.9V\right].$$

Figure 5.10 shows the voltage transfer characteristic determined using the equations above.

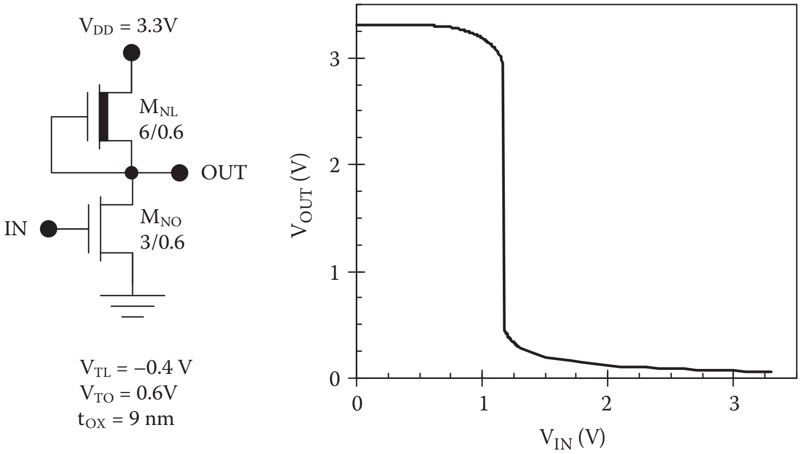


FIGURE 5.10
Calculated VTC for the MOS inverter of Figure 5.9.

Example 5.3 Critical Voltages

Determine the critical voltages of the transfer function for the NMOS inverter shown in Figure 5.11.

Solution: The process transconductance parameter for the transistors is

$$k' = \frac{\mu_n \varepsilon_{OX}}{t_{OX}} = \frac{\left(580cm^2 / Vs\right)(3.9)\left(8.85 \times 10^{-14} F / cm\right)}{9 \times 10^{-7} cm} = 0.22mA / V^2 \cdot$$

The device transconductance parameters are

$$K_L = k' \frac{W_L}{L_L} = \left(0.22mA / V^2\right)\left(\frac{6\mu m}{0.6\mu m}\right) = 2.2mA / V^2$$

and

$$K_O = k' \frac{W_O}{L_O} = \left(0.22mA / V^2\right)\left(\frac{3\mu m}{0.6\mu m}\right) = 1.1mA / V^2 .$$

The output voltage levels are

$$V_{OL} = V_{DD} - V_{TO} \pm \sqrt{\left(V_{DD} - V_{TO}\right)^2 - \left(\frac{K_L}{K_O}\right) V_{TL}^2}$$

$$= 3.3V - 0.6V - \sqrt{(3.3V - 0.6V)^2 - \left(\frac{2.2mA / V^2}{1.1mA / V^2}\right)\left(-0.4V\right)^2} = 0.15V ,$$

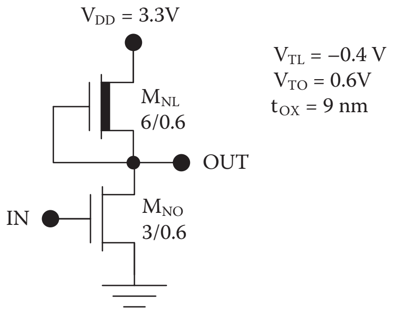


FIGURE 5.11
Example MOS inverter for the calculation of the voltage transfer characteristic.

and

$$V_{OH} = V_{DD} = 3.3V \cdot$$

The critical input voltages are

$$V_{IL} = V_{TO} + \frac{K_L}{\sqrt{K_O K_L + K_O^2}} \left| V_{TL} \right|$$

$$= 0.6V + \frac{2.2mA/V^2}{\sqrt{\left(1.1mA/V^2\right)\left(2.2mA/V^2\right) + \left(1.1mA/V^2\right)^2}} \left| -0.4V \right| = 1.06V \, ,$$

$$V_M = V_{TO} - V_{TL}\sqrt{\frac{K_L}{K_O}} = 0.6V - \left(-0.4V\right)\sqrt{\frac{2.2mA/V^2}{1.1mA/V^2}} = 1.16V \, ,$$

and

$$V_{IH} = V_{TO} + 2\left| V_{TL} \right| \sqrt{\frac{K_L}{3K_O}} = 0.6V + 2\left| -0.4V \right| \sqrt{\frac{2.2mA/V^2}{3\left(1.1mA/V^2\right)}} = 1.25V \, .$$

The ideal value of the switching threshold V_M is $V_{DD}/2$, but because of other design constraints, this may be difficult to achieve in a MOS inverter with a depletion-type load.

5.2 Dissipation

The depletion-loaded MOS inverter draws a steady DC current under the output low condition so both the DC and dynamic dissipation (described in Chapter 1) may be important. 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 5.12 (the input is grounded). With a logic one output, the switch transistor is operating in cutoff and the supply current I_{DDH} is approximately zero (apart from the small leakage current). Therefore,

$$P_H = V_{DD} I_{DDH} \approx 0 \,. \tag{5.28}$$

With a logic zero output as shown in Figure 5.13, the load transistor M_{NL} is saturated while the switch transistor M_{NO} is linear. Therefore, the output low supply current is

$$I_{DDL} = \frac{K_L V_{TL}^2}{2} \,, \tag{5.29}$$

and the static dissipation with the output low is

$$P_L = V_{DD} I_{DDL} = \frac{K_L V_{DD} V_{TL}^2}{2} \,. \tag{5.30}$$

The average static dissipation depends on the output duty cycle, but it is common practice to assume a 50% duty cycle. For this case, the average static dissipation is

$$P_{DC} \approx \frac{P_H + P_L}{2} = \frac{K_L V_{DD} V_{TL}^2}{4} \,. \tag{5.31}$$

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 to preserve the desired voltage transfer characteristic.) Scaling down the K values

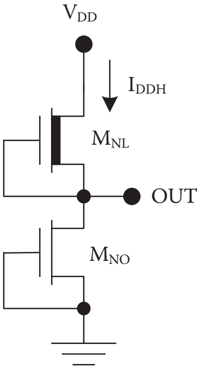


FIGURE 5.12
MOS inverter for the determination of P_H.

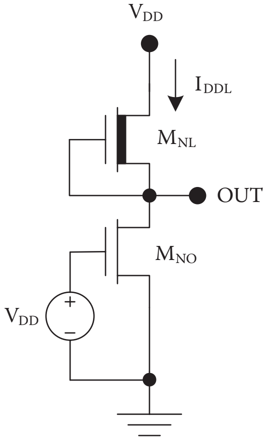


FIGURE 5.13
MOS inverter for the determination of P_L.

degrades the switching speed of the circuitry unless the load capacitances can be scaled similarly.

The dynamic or AC dissipation is associated with the charging and discharging of the load capacitance (the capacitance switching power). Consider the MOS inverter with a lumped capacitive load as shown in Figure 5.14.

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

$$J = V_{DD} \int_{\substack{clock\\cycle}} i_{DD} dt \,. \tag{5.32}$$

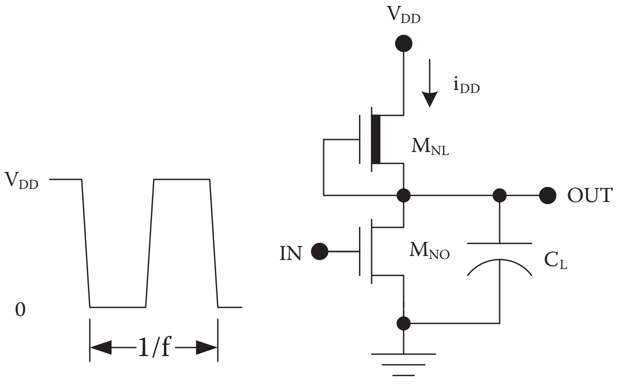


FIGURE 5.14
MOS inverter for the consideration of the dynamic dissipation.

If we neglect the current that flows in M_{NO} during the low-to-high transition (the *crossover current*), then

$$i_{DD} = C_L \frac{dV_{OUT}}{dt}, \tag{5.33}$$

so that

$$J = V_{DD} \int_0^{V_{DD}} C_L dV_{OUT} = C_L V_{DD}^2 . \tag{5.34}$$

This is the energy, in joules, dissipated during each switching cycle. The capacitance switching dissipation is therefore

$$P_{switch} = fC_L V_{DD}^2 = \alpha f_{CLK} C_L V_{DD}^2 , \tag{5.35}$$

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 that 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 capacitance switching components:

$$P = P_{DC} + P_{switch} = \frac{K_L V_{DD} V_{TL}^2}{4} + fC_L V_{DD}^2 . \tag{5.36}$$

Example 5.4 MOS Inverter Dissipation

For the MOS inverter of Figure 5.15, calculate the dissipation as a function of the switching frequency.

Solution: The static dissipation is

$$P_{DC} \approx \frac{P_H + P_L}{2} = \frac{K_L V_{DD} V_{TL}^2}{4} = \frac{(2.2mA/V^2)(3.3V)(-0.4V)^2}{4} = 0.29mW .$$

If we neglect the crossover current associated with simultaneous conduction of the pull-down and pull-up transistors, the dynamic (capacitance switching) dissipation is given by

$$P_{switch} = fC_L V_{DD}^2 = f(5pF)(3.3V)^2 = f(54pJ) .$$

The total dissipation is

$$P = P_{DC} + P_{switch} = 0.29mW + f(54pJ) .$$

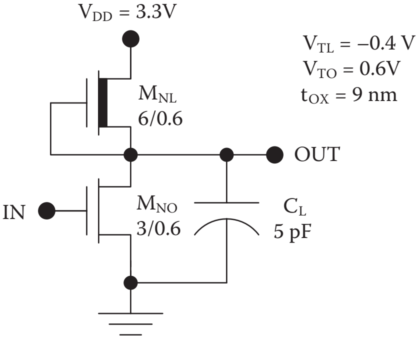


FIGURE 5.15
MOS inverter for the calculation of the dissipation.

The dynamic and static dissipation components will be equal at a switching frequency of 5.4 MHz, but for lower frequencies, the static dissipation is dominant as shown in Figure 5.16.

5.4 Propagation Delays

To estimate the propagation delays of an MOS inverter with a depletion type pull up, we will consider the loading to be a lumped capacitance connected between the output and ground as shown in Figure 5.17. Real loads involve

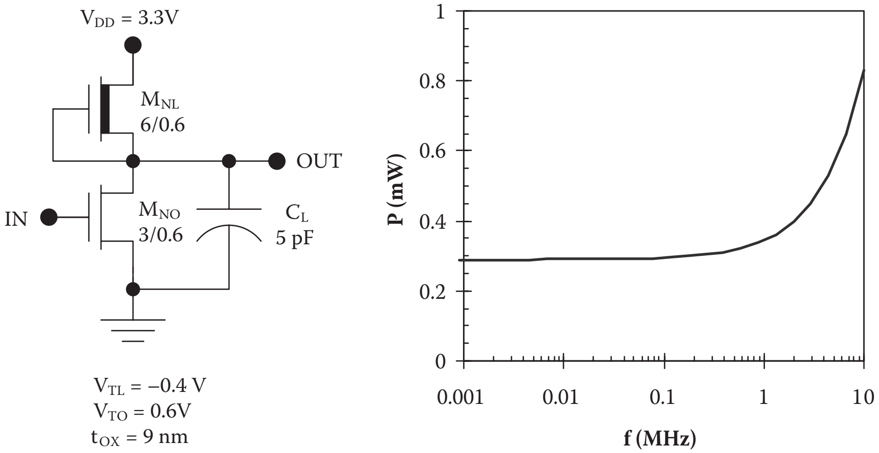


FIGURE 5.16
Example calculation of the dissipation versus switching frequency for an MOS inverter.

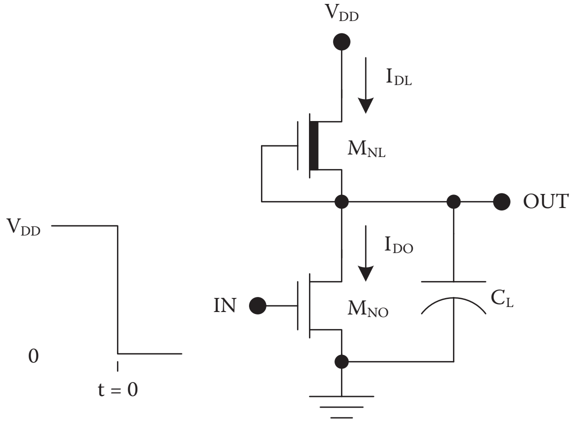


FIGURE 5.17
MOS inverter for the calculation of t_{PLH}.

distributed capacitances, resistances, and inductances, so the lumped capacitive load is an approximation.

To estimate the low-to-high propagation delay t_{PLH}, we will assume that the fall time at the input is negligible. (Whereas 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} so that $I_{DO} \approx 0$. M_{NL} is saturated, and the drain current is given by

$$I_{DL} = \frac{K_L V_{TL}^2}{2}. \tag{5.37}$$

Therefore, the time derivative of the output voltage is

$$\frac{dV_{OUT}}{dt} = \frac{1}{C_L}\frac{K_L V_{TL}^2}{2}. \tag{5.38}$$

By definition of the propagation delay, V_{OUT} reaches $V_{DD}/2$ at $t = t_{PLH}$. Solving,

$$t_{PLH} = \frac{V_{DD} C_L}{K_L V_{TL}^2}. \tag{5.39}$$

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 also be estimated simply by assuming that the load is a lumped capacitance and that the rise time at the input is negligible (see Figure 5.18).

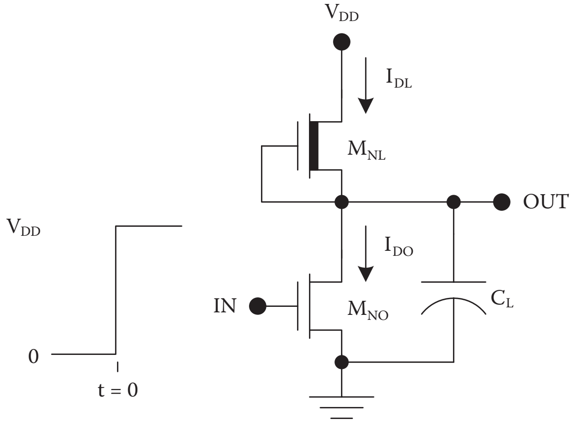


FIGURE 5.18
MOS inverter for the estimation of t_{PHL}.

At t = 0^+, M_{NO} will be saturated and M_{NL} will be linear. Thus, at t = 0^+, the drain currents are given by

$$I_{DO} = \frac{K_O}{2}\left(V_{DD} - V_{TO}\right)^2 \tag{5.40}$$

and

$$I_{DL} = K_L\left[\left|V_{TL}\right|\left(V_{DD} - V_{OUT}\right) - \frac{\left(V_{DD} - V_{OUT}\right)^2}{2}\right]. \tag{5.41}$$

The drain current in the load transistor depends on the square of the output voltage, resulting in a nonlinear differential equation. However, we can greatly simplify the analysis by neglecting I_{DL}. By neglecting the simultaneous conduction of the two transistors (the crossover current), the error is usually less than 20%, and this is acceptable for hand calculations. Within this approximation, we obtain

$$t_{PHL} \approx \frac{V_{DD}C_L}{K_O\left(V_{DD} - V_{TO}\right)^2}. \tag{5.42}$$

Therefore, 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 5.5 MOS Inverter Propagation Delays

Estimate the propagation delays for the NMOS inverter shown in Figure 5.19.

Solution: The low-to-high propagation delay is

$$t_{PLH} = \frac{V_{DD}C_L}{K_L V_{TL}^2} = \frac{(3.3V)(5pF)}{(2.2mA/V^2)(-0.4V)^2} = 47ns\ .$$

The high-to-low propagation delay is

$$t_{PHL} \approx \frac{V_{DD}C_L}{K_O\left(V_{DD} - V_{TO}\right)^2} = \frac{(3.3V)(5pF)}{(1.1mA/V^2)(3.3V - 0.6V)^2} = 2.1ns\ .$$

Therefore, t_{PLH} is ~20 times longer than t_{PHL}! To make the propagation delays more nearly equal, we would need to either increase K_L/K_O or make V_{TL} more negative. Either modification would increase V_{OL} so that these choices would have to be made with due consideration of the DC voltage transfer characteristic.

5.5 Fan-Out

The fan-out of MOS gates is determined by dynamic rather than DC (static) considerations. This is because the load gates (assumed to be similar MOS gates) present primarily capacitive loading. If the loading associated with internal capacitances (in the gate circuit) and interconnect (because of wires between the gate circuits) may be neglected, then the load capacitance increases in direct proportion to the number of fan-out gates. Inasmuch as

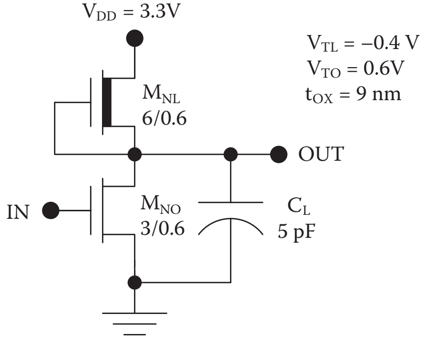


FIGURE 5.19
Example MOS inverter for the calculation of the propagation delays.

the propagation delay is proportional to the fan-out, the maximum fan-out is dictated by the maximum tolerable propagation delay.*

Consider an MOS inverter loaded by N similar circuits as shown in Figure 5.20. If a ceiling has been established for the propagation delay, then the maximum allowable load capacitance can be determined from the worst-case propagation delay:

$$C_{L,\max} = \min\left(\frac{K_L V_{TL}^2}{V_{DD}} t_{p,\max}\ ,\ \frac{K_O\left(V_{DD} - V_{TO}\right)^2}{V_{DD}} t_{p,\max}\right), \tag{5.43}$$

and the maximum fan-out is the largest integer satisfying

$$N \le \frac{C_{L,\max}}{C_{in}}, \tag{5.44}$$

where the input capacitance for one fan-out gate may be estimated as the worst-case oxide capacitance for the pull-down transistor:

$$C_{in} = W_O\left(L_O + 2L_{OV}\right)C_{ox}. \tag{5.45}$$

Simple hand calculations made with these approximations are expected to yield better than factor-of-two accuracy. More precise estimates should take into account the bias dependence of C_{gs} and C_{gd} in the pull-down transistors of the fan-out gates as well as the loading by C_{gd} and C_{sb} in the pull-up transistors of the load gates. Internal loading attributable to capacitances in the driving gate may also have to be considered if N is small.

Example 5.6 Fan-Out for MOS Inverters

Estimate the maximum fan-out for the MOS inverter design illustrated in Figure 5.21 if (with a system clock frequency of 500 MHz) the maximum allowable propagation delay is 100 ps.

Solution: The device transconductance parameters are

$$K_L = \frac{W_L}{L_L}\frac{\mu_n \varepsilon_{ox}}{t_{ox}} = \left(\frac{6\mu m}{0.6\mu m}\right)\left(\frac{\left(580cm^2/Vs\right)(3.9)\left(8.85\times10^{-14}F/cm\right)}{90\times10^{-8}cm}\right.$$

$$= 2.2mA/V^2$$

* The maximum propagation delay is inversely proportional to the system clock frequency $t_{p,max} = K/f$. However, the constant of proportionality K depends on the system architecture (for example, the number of stages a signal must ripple through in one clock period) so that its determination is quite complex.

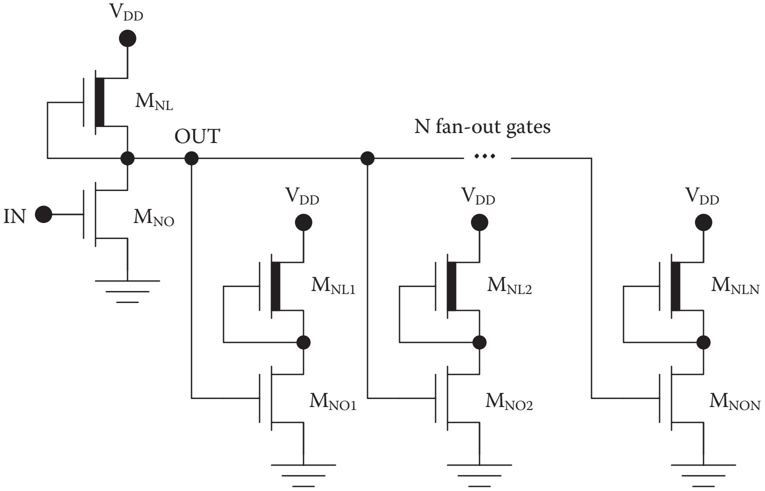


FIGURE 5.20
MOS inverter with N similar fan-out gates.

and

$$K_O = \frac{W_O}{L_O}\frac{\mu_n \varepsilon_{ox}}{t_{ox}} = \left(\frac{3\mu m}{0.6\mu m}\right)\left(\frac{\left(580cm^2/Vs\right)(3.9)\left(8.85\times10^{-14}F/cm\right)}{90\times10^{-8}cm}\right)$$

$$= 1.1mA/V^2.$$

The maximum allowable load capacitance is

$$C_{L,\max} = \min\left(\frac{K_L V_{TL}^2}{V_{DD}}t_{p,\max}\ ,\ \frac{K_O\left(V_{DD}-V_{TO}\right)^2}{V_{DD}}t_{p,\max}\right)$$

$$= \min\left(\frac{\left(2.2mA/V^2\right)(-0.4V)^2}{3.3V}100ps,\ \frac{\left(1.1mA/V^2\right)(3.3V-0.6V)^2}{3.3V}100ps\right)$$

$$= \min\left(0.107pF\ ,\ 2.43pF\right) = 0.107pF.$$

The gate oxide capacitance per unit area for the MOSFETs is

$$C_{ox} = \frac{\varepsilon_{ox}}{t_{ox}} = \frac{(3.9)\left(8.85\times10^{-14}F/cm\right)}{90\times10^{-8}cm} = 3.8\times10^{-7}F/cm^2,$$

and the approximate input capacitance for each fan-out gate is

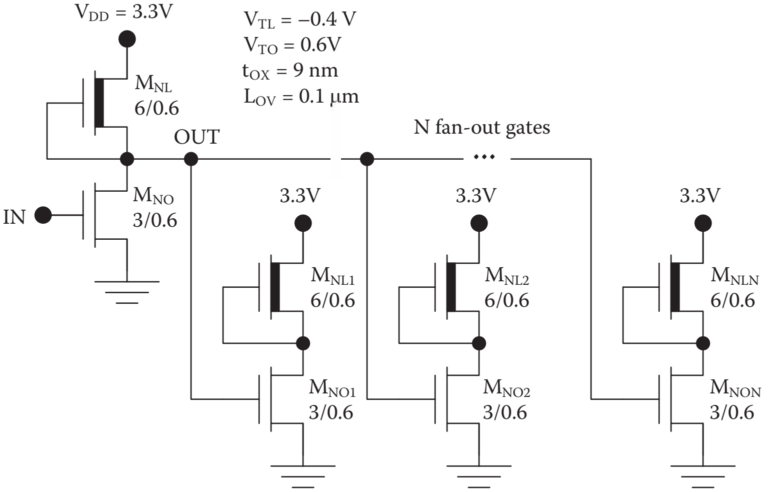


FIGURE 5.21
MOS inverter fan-out calculation example.

$$\begin{aligned} C_{in} &= W_O\left(L_O + 2L_{OV}\right)C_{ox} \\ &= 3\times10^{-4}cm\left(0.6\times10^{-4}cm + 2\left(0.1\times10^{-4}cm\right)\right)\left(3.8\times10^{-7}F/cm^2\right) \\ &= 9.1fF. \end{aligned}$$

The maximum fan-out is the largest integer satisfying

$$N \le \frac{C_{L,\max}}{C_{in}} = \frac{0.107pF}{9.1fF} = 11.7\,,$$

so N = 11.

5.6 NOR Circuits

An MOS NOR gate may be realized by placing pull-down transistors in parallel, as shown in Figure 5.22 for the case of three inputs.

Here, if the voltage representing logic one is applied to one or more of the pull-down transistors, the output will go low; otherwise, the output will go high. If the transistors are designed identically to those in an inverter circuit, then the worst-case electrical characteristics will be approximately the same as for the inverter and all of the same equations may be used for approximate

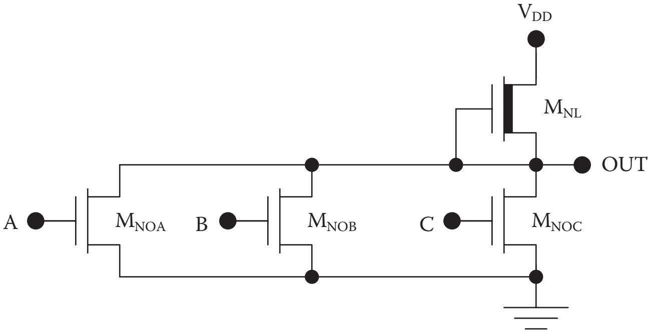


FIGURE 5.22
MOS three-way NOR gate.

hand analysis. By worst case, we mean the situation in which only one pull-down transistor conducts. The simultaneous conduction of two or more pull-down devices will improve both V_{OL} and t_{PHL} compared with the reference inverter circuit.

Although the inverter equations apply approximately for worst-case analysis, the NOR circuit does have additional capacitive loading of the output node, which is associated with the added pull-down transistors. Typically, we should be able to neglect this unless the number of inputs (fan-in) is comparable with the fan-out.

5.7 NAND Circuits

An MOS NAND gate is made by placing the pull-down transistors in series; this is shown in Figure 5.23 for the case of three inputs.

Here the output goes low if a logic one voltage is applied to all inputs, causing all pull-down transistors to operate in the linear region. Otherwise, the output goes high. For an M-way NAND gate, it is necessary to scale up the widths of the pull-down transistors by a factor of approximately M to achieve electrical characteristics comparable with the reference inverter. This is because the pull-down transistors conduct in series. For example, to maintain the same value of V_{OL} as in the reference inverter, each pull-down transistor should have a drain-to-source voltage of ~V_{OL}/M. This indicates that the pull-down transistors must be M times as wide as those in the reference inverter (an inverter having otherwise same electrical characteristics). The same argument may be made with the high-to-low propagation delay and the fall time, which will be governed by the current-sinking ability of the M pull-down devices connected in series.

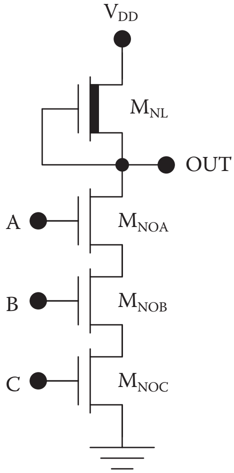


FIGURE 5.23
Three-way MOS NAND gate.

Two factors place a practical limit on the fan-in for an MOS NAND circuit. 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 M^2). Second, both the static and dynamic performance will be degraded even if the widths of the pull-down 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 therefore will have higher on-resistances and larger values of V_{DS}, therefore degrading V_{OL} and t_{PHL}. Moreover, because the p-type bodies of the pull-down transistors are tied to ground, only the bottom transistor has zero body-to-source bias. The other pull-down transistors will experience non-zero (but different) body-to-source bias voltages that will modify their threshold voltages and further complicate the analysis. Because of the increased circuit area and inferior performance of NAND gates, NOR circuits are preferred when other factors are equal.

5.8 Exclusive OR (XOR) Circuit

The exclusive OR function can be implemented in NMOS using the ingenious circuit of Figure 5.24, which operates 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} in cutoff, 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} in

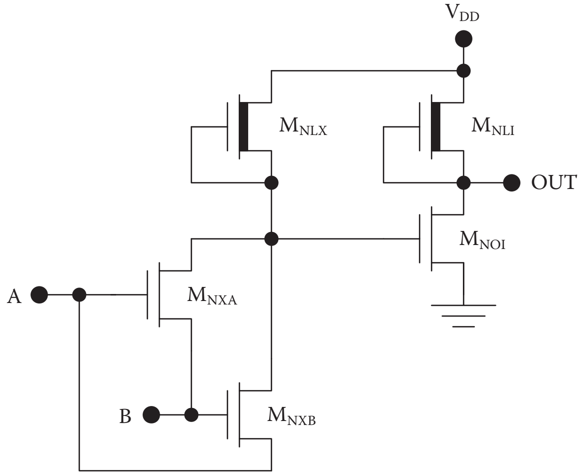


FIGURE 5.24
MOS two-way XOR gate.

cutoff, the output goes high as before. On the other hand, if the inputs are the same (both logic zero or both logic one), then both M_{NXA} and M_{NXB} are in cutoff, the voltage at the gate of M_{NOI} goes high, and the output voltage goes low.

5.9 General Logic Design

General AND-OR-INVERT functions may be implemented by combining parallel and series connected pull-down transistors. In the example circuit of Figure 5.25, the inputs A, B, and C are ANDed by the series combination of the transistors M_{NOA}, M_{NOB}, and M_{NOC}. In similar manner, inputs E and F are ANDed. These two results are ORed with input D by the parallel combination of the three circuit branches, so that the overall logic function is

$$Y = \overline{ABC + D + EF}\,. \tag{5.39}$$

Another example circuit is shown in Figure 5.26. Here, the overall logic function is

$$Y = \overline{\left[(A+B)C(D+E+F)\right] + \left[(G+H+I)(J+K)\right]}\,. \tag{5.40}$$

For the implementation of a general logic function in NMOS, the pull-down transistor must be scaled in width by a factor R_i compared with the transistors in the reference inverter. The scale factor, which may not be the same

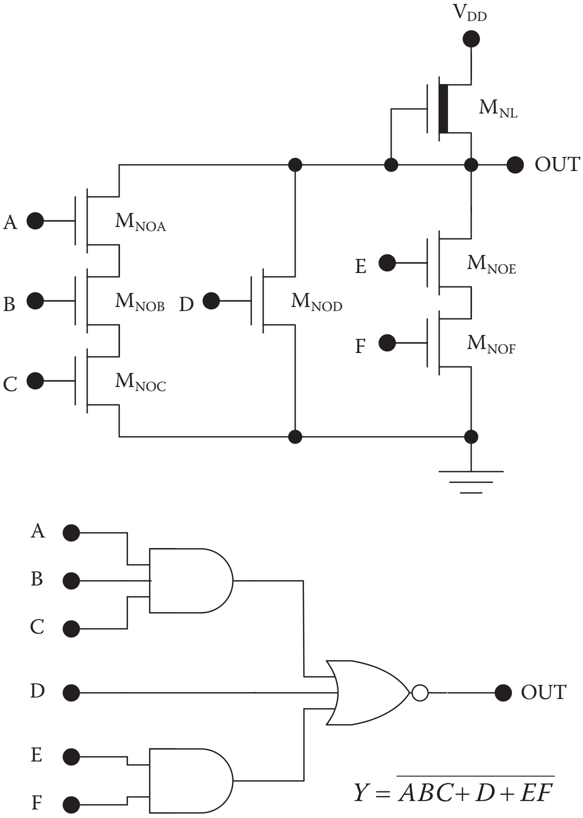
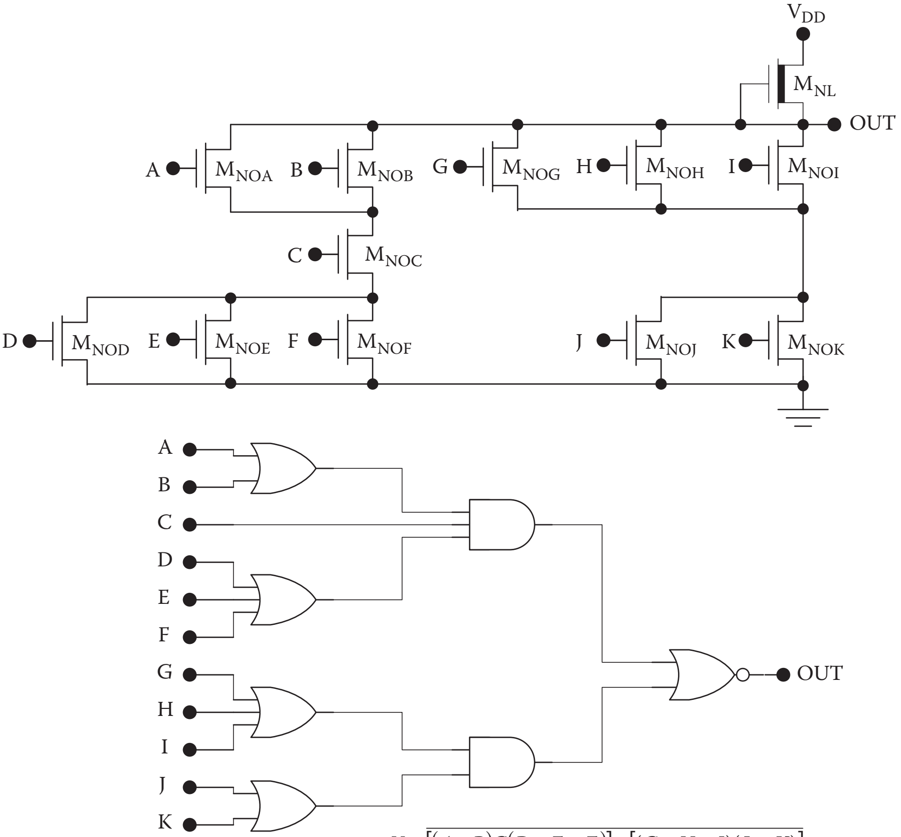


FIGURE 5.25
An AND-OR-INVERT circuit with six inputs.

for all transistors, is equal to the maximum number of series transistors for all paths connecting the output node to ground and containing transistor i. In the circuit of Figure 5.26, the switch transistors in the left branch (M_{NOA}, M_{NOB}, M_{NOC}, M_{NOD}, M_{NOE}, and M_{NOF}) should be scaled by a factor of three compared with the reference inverter, whereas the switch transistors in the right branch (M_{NOG}, M_{NOH}, M_{NOI}, M_{NOJ}, and M_{NOK}) should be scaled by a factor of two. In practice, this scaling is achieved by increasing the gate widths while keeping the gate lengths fixed.

5.10 Pass Transistor Circuits

Up to now, we have considered logic circuits in which all inputs all applied to the gates of MOS transistors. Occasionally, it is useful to connect inputs to



$$Y = \overline{[(A+B)C(D+E+F)]+[(G+H+I)(J+K)]}$$

FIGURE 5.26
An AND-OR-INVERT circuit with 11 inputs.

the S/D terminals of MOSFETs. A single n-MOS transistor can be used as a pass transistor as shown in Figure 5.27; here the input voltage is transferred to the output if the pass transistor is enabled by applying a positive voltage at the gate, causing linear operation of the MOSFET. If the enable signal is brought to zero (or the most negative voltage in the circuit), the n-MOS transistor will be cutoff and no current will flow between the input and output in this *high impedance* state. Pass transistors play important roles in dynamic circuits (discussed in Chapter 8) and memory circuits (discussed in Chapter 11).

An extension of the pass transistor is the CMOS transmission gate, shown in Figure 5.28, which combines n-MOS and p-MOS transistors to achieve low "on" resistance over the entire range of input voltages from 0 to V_{DD}. Transmission gates are described in more detail in Chapter 12.

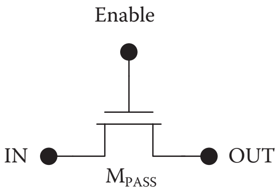


FIGURE 5.27
n-MOS pass transistor.

It is also possible to use pass transistors to create Boolean logic gates. For example, an AND gate may be constructed as shown in Figure 5.29. In this circuit, if input B is logic "0", the output of the inverter will go high, M_{PD} will be linear, and the output will go low. If both inputs A and B go high, M_1 will be linear, M_{PD} will be cutoff, and the output will go high.

Other, more complicated logic functions can be implemented in pass transistor logic and its variations. Advantages are potential savings in silicon area, by using fewer transistors for a particular logic function, and a reduction in switching energy, attributable to reduced logic swing.

5.11 SPICE Demonstrations

For the purpose of illustration, simulations were performed using Cadence Capture CIS 10.1.0 PSpice (Cadence Design Systems). The level 1 MOS transistor model parameters given in Tables 5.1 and 5.2 were used unless otherwise

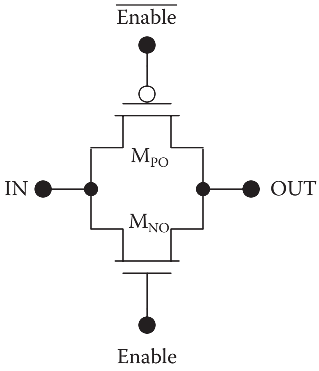


FIGURE 5.28
CMOS transmission gate.

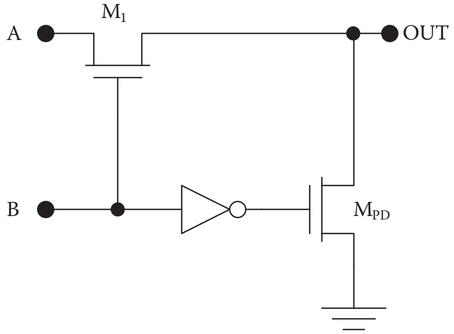


FIGURE 5.29
Pass transistor logic two-way AND gate.

noted. The process transconductance parameters were calculated assuming an oxide thickness of 9 nm. For n-MOSFETS,

$$KP = \frac{(3.9)\left(8.85\times10^{-14} F/cm\right)\left(580 cm^2V^{-1}s^{-1}\right)}{9\times10^{-7} cm} = 222\mu A/V^2, \qquad (5.46)$$

and for p-MOSFETS,

$$KP = \frac{(3.9)\left(8.85\times10^{-14} F/cm\right)\left(230 cm^2V^{-1}s^{-1}\right)}{9\times10^{-7} cm} = 88\mu A/V^2, \qquad (5.47)$$

The overlap capacitances per unit gate width were determined with the assumption that $L_{OV} = 0.1\mu m$:

TABLE 5.1
n-MOS Level 1 SPICE Parameters

Parameter	Value	Units
KP	222u	A/V^2
VTO	0.5	V
GAMMA	0.15	V$^{1/2}$
PHI	0.7	V
LAMBDA	0.05	
TOX	9n	m
NSUB	1E16	cm^{-3}
UO	580	cm^2/Vs
CGSO	0.38n	F/m
CGDO	0.38n	F/m

TABLE 5.2

Depletion-Type n-MOS Level 1 SPICE Parameters

Parameter	Value	Units
KP	222u	A/V^2
VTO	–0.3	V
GAMMA	0.15	V$^{1/2}$
PHI	0.7	V
LAMBDA	0.05	
TOX	9n	m
NSUB	1E16	cm^{-3}
UO	580	cm^2/Vs
CGSO	0.38n	F/m
CGDO	0.38n	F/m

$$\begin{aligned} CGSO &= \frac{(3.9)\left(8.85\times10^{-14}F/cm\right)\left(0.1\times10^{-4}cm\right)}{9\times10^{-7}cm} \\ &= 3.8pF/cm = 0.38nF/m \end{aligned} \tag{5.48}$$

and

$$\begin{aligned} CGDO &= \frac{(3.9)\left(8.85\times10^{-14}F/cm\right)\left(0.1\times10^{-4}cm\right)}{9\times10^{-7}cm}. \\ &= 3.8pF/cm = 0.38nF/m \end{aligned} \tag{5.49}$$

The body effect coefficient was calculated from

$$\begin{aligned} GAMMA &= \frac{\sqrt{2q\varepsilon_{Si}N_a}}{C_{ox}} \\ &= \frac{\sqrt{2\left(1.602\times10^{-19}C\right)(11.9)\left(8.85\times10^{-14}F/cm\right)\left(10^{16}cm^{-3}\right)}}{(3.9)\left(8.85\times10^{-14}F/cm\right)/9\times10^{-7}cm}. \\ &\approx 0.15V^{1/2} \end{aligned}$$

SPICE Example 5.1. MOS Inverter Voltage Transfer Characteristic

Using a DC sweep of the input voltage, the transfer characteristic was determined for an MOS inverter with a depletion type load as shown in Figure 5.30. The depletion type transistor has 10 times the width of the pull-down transistor, and the resulting output low voltage is $V_{OL} \approx 0.23V$ as shown by the results in Figure 5.31.

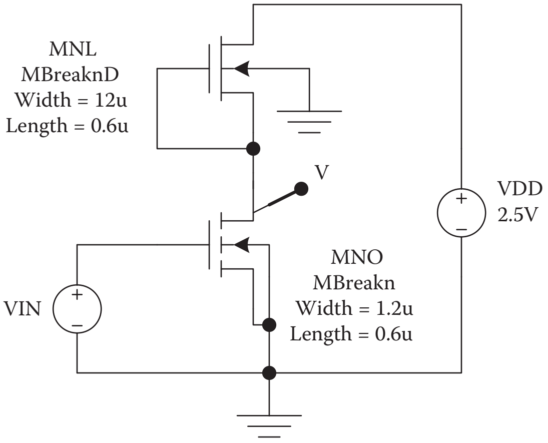


FIGURE 5.30
MOS inverter circuit for the determination of the voltage transfer characteristic.

SPICE Example 5.2 MOS Inverter Propagation Delays

Propagation delays were determined using a transient simulation for the inverter of Figure 5.32. A load capacitance of 1 pF was used, and the pulse source parameters were V1 = 0, V2 = 2.5V, TD = 0, TF = 10 ns, TR = 10 ns, PW = 40 ns, and PER = 100 ns. The propagation delays, determined from the results of Figure 5.33, are $t_{PHL} = 3.8ns$ and $t_{PLH} = 6.8ns$.

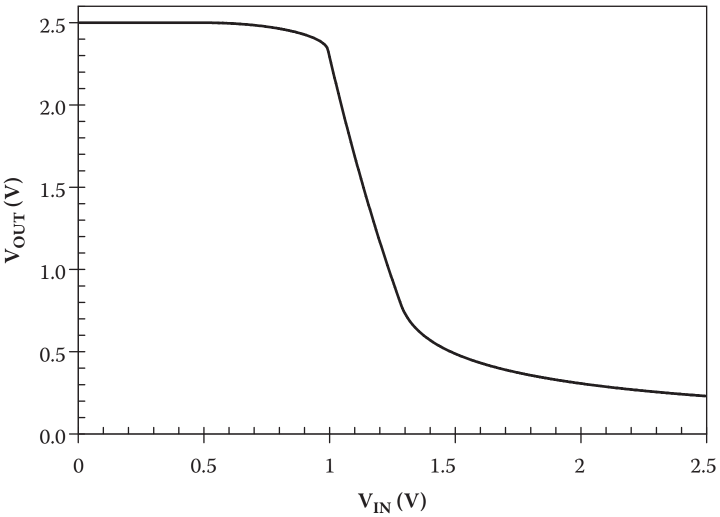


FIGURE 5.31
Voltage transfer characteristic for the MOS inverter of Figure 5.30.

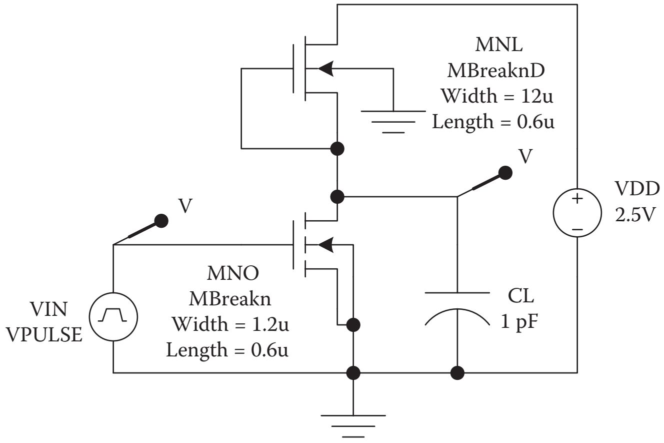


FIGURE 5.32
MOS inverter circuit for the determination of the propagation delays with a 1 pF load.

5.12 Practical Perspective

For practical perspective articles, see the dynamic website at http://www.engr.uconn.edu/ece/books/ayers.

5.13 Summary

MOS logic gates may be implemented using n-MOS transistors in a pull-down logic network along with a passive pull-up device. The pull-up device may be a depletion-type n-MOS transistor. The static voltage transfer characteristic for such a logic circuit can be determined by equating the currents in the pull-up and pull-down networks. For an inverter with abrupt input voltage transitions and a lumped load capacitance, the low-to-high propagation delay is proportional to the load capacitance and inversely proportional to the device transconductance parameter for the pull-up transistor. The high-to-low propagation delay is proportional to the load capacitance and inversely proportional to the device transconductance parameter for the pull-down transistor. NAND gates may be implemented by placing pull-down transistors in series, NOR gates may be implemented by placing pull-down transistors in parallel, and other more complex logic functions may be implemented by series and parallel combinations of switch transistors. MOS logic gates of this type do not require p-MOS transistors and are convenient

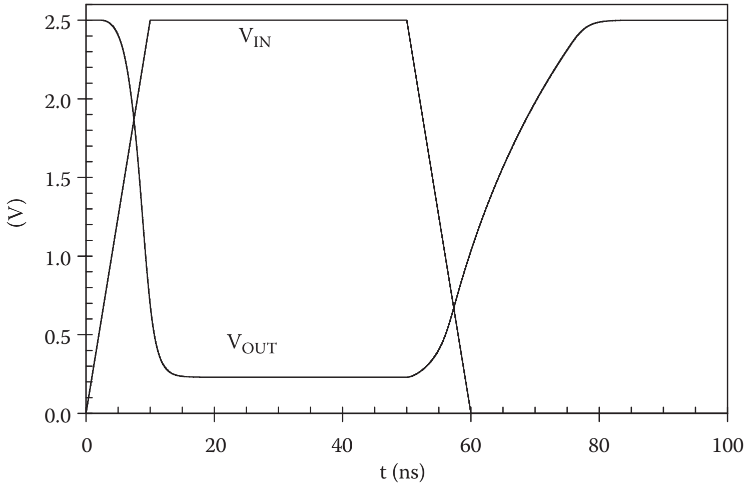


FIGURE 5.33
Transient response for an MOS inverter with a 1 pF load.

for use in some memory circuits. However, a significant drawback with this type of circuit is the static dissipation.

5.14 Exercises

E5.1. For the MOS inverter of Figure 5.34, (1) determine V_{IL}, V_{IH}, V_{OL}, and V_{OH}, and (2) determine the noise margins V_{NML} and V_{NMH}.

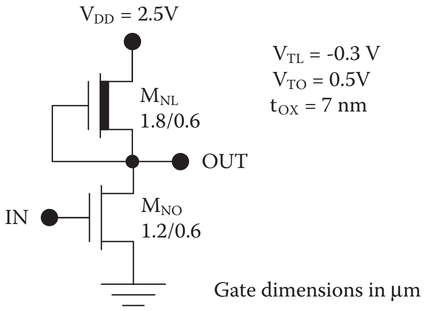


FIGURE 5.34
MOS inverter for the determination of the critical voltages (see Exercise E5.1).

E5.2. For the MOS inverter of Figure 5.35, calculate and plot the static voltage transfer characteristic. From the numerical values, determine V_{IL} and V_{IH}.

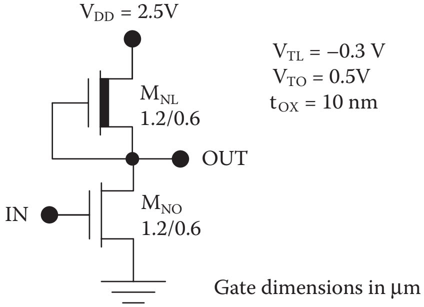


FIGURE 5.35
MOS inverter for the determination of the voltage transfer characteristic (see Exercise E5.2).

E5.3 For the MOS inverter of Figure 5.36, find the ranges of V_{IN} for cutoff, saturated, and linear operation of M_{NO}.

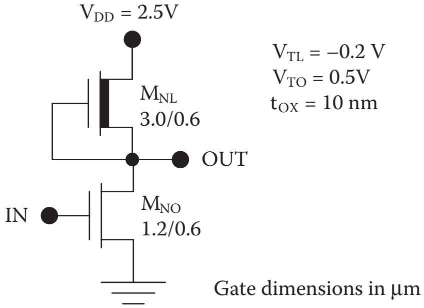


FIGURE 5.36
MOS inverter for the analysis of the static behavior (see Exercise E5.3).

E5.4 For the MOS inverter of Figure 5.37, determine the average DC dissipation and find the maximum packing density in gates per square centimeter if the maximum power density is 10 W/cm^2.

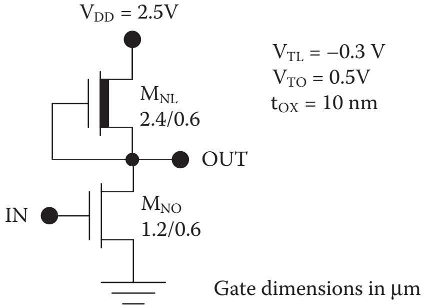


FIGURE 5.37
MOS inverter for the consideration of dissipation (see Exercise E5.4).

E5.5. Determine the propagation delays for the inverter of Figure 5.38.

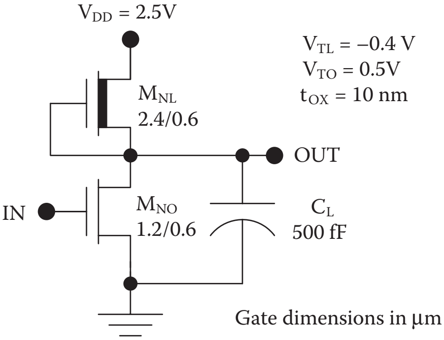


FIGURE 5.38
MOS inverter for the determination of the propagation delays (see Exercise E5.5).

E5.6 The inverter shown in Figure 5.39 is loaded by five identical inverters. Determine the propagation delays.

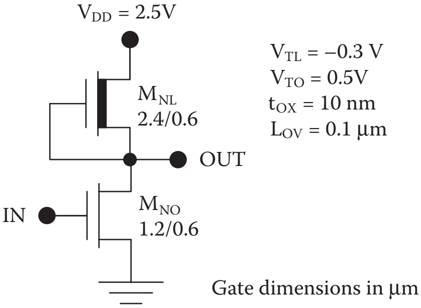


FIGURE 5.39
MOS inverter for the determination of the propagation delays (see Exercise E5.6).

E5.7 Suppose a ring oscillator is constructed using inverters with the design shown in Figure 5.40. How many stages will result in an oscillation frequency of 10 MHz? Estimate the total power dissipation for this M-stage ring while it is oscillating.

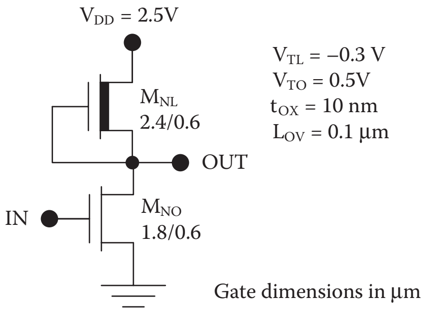


FIGURE 5.40
MOS inverter for the construction of a ring oscillator (see Exercise E5.7).

E5.8 Create the layout design for MOS inverter using a depletion-type transistor with $V_{TL} = -0.3V$ and an enhancement-type device with $V_{TO} = 0.5V$ such that $t_{PLH} \le 1ns$ and $t_{PHL} \le 1ns$ with $C_L = 1pF$. $V_{DD} = 2.5V$, $t_{ox} = 10nm$, and $2X = 700nm$.

E5.9 Create the layout design for MOS inverter using a depletion-type transistor with $V_{TL} = -0.3V$ and an enhancement-type device with $V_{TO} = 0.5V$ such that $V_{OL} \le 0.1V$. $V_{DD} = 2.5V$, $t_{ox} = 10nm$, and $2X = 0.6\mu m$.

E5.10 Create the layout design for MOS NAND3 gate using a depletion-type transistor with $V_{TL} = -0.3V$ and enhancement-type transistors with $V_{TO} = 0.5V$ such that $V_{OL} \le 0.1V$. $V_{DD} = 2.5V$, $t_{ox} = 10nm$, and $2X = 0.6\mu m$.

For additional exercise problems, see the dynamic website at http://www.engr.uconn.edu/ece/books/ayers.

6

Static CMOS

6.1 Introduction

CMOS logic, which uses complementary n-channel and p-channel MOS transistors, is by far the most important logic family today because of its low standby power, high packing density, and high speed. CMOS is used extensively in high-performance, portable (battery-operated) products such as notebook computers and wireless handheld devices.

Figure 6.1 shows some basic logic gates with their CMOS circuit realizations. On the left, these fundamental logic circuits are shown using general active-high and active-low switches. On the right, the same logic gates have been realized in CMOS, using n-MOS transistors and p-MOS transistors for the active-high and active-low switches, respectively. The inverter requires one n-MOS transistor and one p-MOS transistor, whereas the two-way NAND and NOR gates require two of each type of transistor. The total number of transistors required is two times the number of inputs in all cases.

CMOS gates use all enhancement-type transistors for low standby dissipation. In the logic one output state, the transistors of the pull-up branch are conducting but those of the pull-down branch are not. On the other hand, for the logic zero output state, only the pull-down branch is conducting. The only simultaneous conduction (accompanied by a *crossover current* from V_{DD} to ground) occurs during switching transitions.

The sections immediately after this will give detailed descriptions of the electrical characteristics for the inverter, and later sections will extend these descriptions to NAND, NOR, and AND-OR-INVERT circuits.

6.2 Voltage Transfer Characteristic

The voltage transfer characteristic for the CMOS inverter of Figure 6.2 may be determined point-by-point by equating the drain currents in the two transistors.*

* Here we adopt the convention that the drain current of the n-MOS transistor flows into the drain, but the drain current of the p-MOS transistor flows out of the drain.

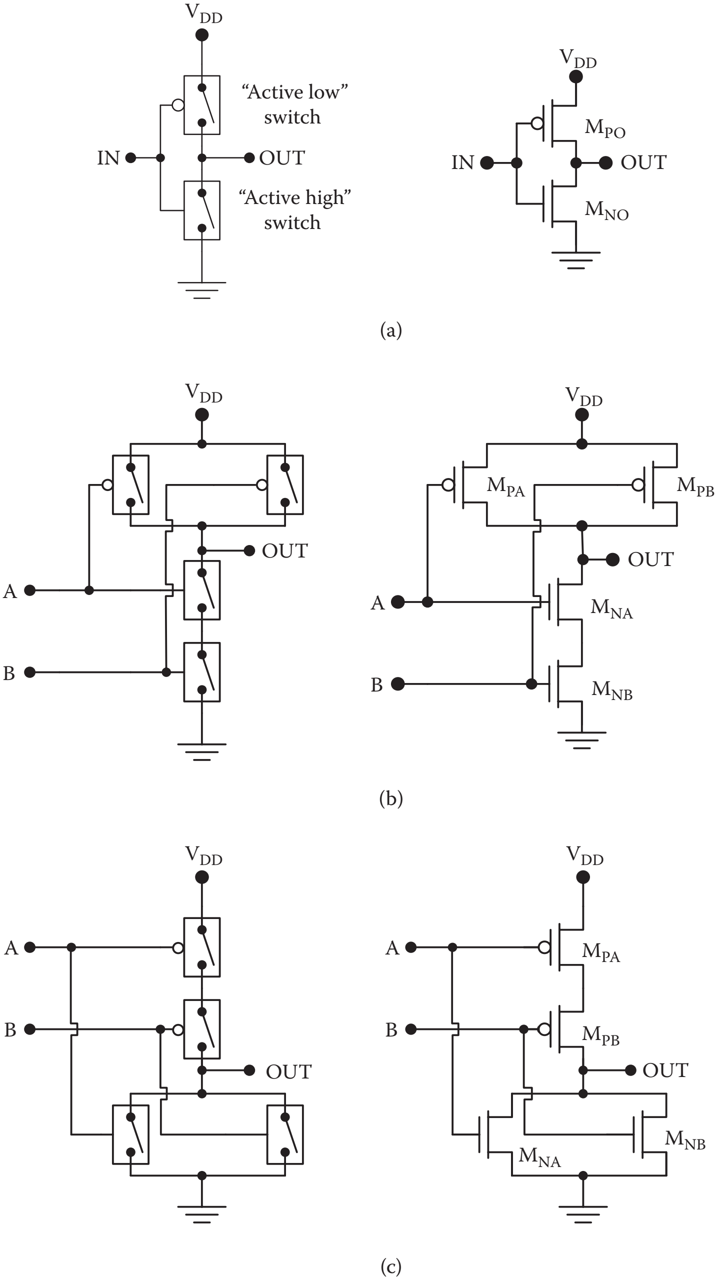


FIGURE 6.1
Basic logic circuits, showing their switch representation and the CMOS realization: (a) Inverter, (b) two-way NAND gate, and (c) two-way NOR gate.

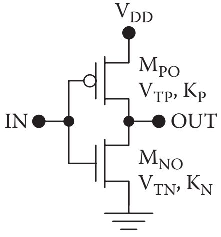


FIGURE 6.2
CMOS inverter.

For this analysis, it is necessary to determine the modes of operation for the two transistors based on their bias voltages as summarized in Table 6.1. The n-MOS transistor is cutoff if $V_{GS} \leq V_{TN}$; in terms of the circuit voltages, this condition is $V_{IN} \leq V_{TN}$, where V_{TN} is the threshold voltage. When conducting, the n-MOS transistor is saturated if $V_{DS} \geq (V_{GS} - V_{TN})$, or, in terms of the circuit voltages, $V_{OUT} \geq (V_{IN} - V_{TN})$; otherwise, the n-MOSFET is linear. For the p-MOS transistor, all of the voltages change signs and the inequalities change direction. Thus, the p-MOS transistor is cutoff if $V_{GS} \geq V_{TP}$ or $V_{IN} \geq (V_{DD} + V_{TP})$, where V_{TP} is the (negative) threshold voltage. The p-MOS device is saturated if $V_{DS} \leq (V_{GS} - V_{TP})$ or $V_{OUT} \leq (V_{IN} - V_{TP})$, but otherwise it is linear.

Consideration of the voltage-based rules for determining the modes of operation for the two devices reveals that there are five regimes in the voltage transfer characteristic, as shown in Table 6.2. These will now be

TABLE 6.1
Voltage Conditions for the Modes of Operation of the n-MOS and p-MOS Transistors in a CMOS Inverter

	n-MOS	p-MOS
Mode	**Voltage conditions**	**Voltage conditions**
Cutoff	$V_{GS} \leq V_{TN}$	$V_{GS} \geq V_{TP}$
	$V_{IN} \leq V_{TN}$	$V_{IN} \geq (V_{DD} + V_{TP})$
Saturation	$V_{DS} \geq (V_{GS} - V_{TN})$	$V_{DS} \leq (V_{GS} - V_{TP})$
	$V_{OUT} \geq (V_{IN} - V_{TN})$	$V_{OUT} \leq (V_{IN} - V_{TP})$
Linear	$V_{DS} \leq (V_{GS} - V_{TN})$	$V_{DS} \geq (V_{GS} - V_{TP})$
	$V_{OUT} \leq (V_{IN} - V_{TN})$	$V_{OUT} \geq (V_{IN} - V_{TP})$

TABLE 6.2

Regimes of Static Operation for the CMOS Inverter

Regime	Voltage conditions	n-MOS mode	p-MOS mode
1	$V_{IN} \le V_{TN}$	Cutoff	Linear
2	$V_{TN} \le V_{IN} \le (V_{OUT} - V_{TN})$	Saturated	Linear
3	$V_{TP} \le (V_{IN} - V_{OUT}) \le V_{TN}$	Saturated	Saturated
4	$(V_{OUT} + V_{TP}) \le V_{IN} \le (V_{DD} + V_{TP})$	Linear	Saturated
5	$V_{IN} \ge (V_{DD} + V_{TP})$	Linear	Cutoff

considered in detail to determine the analytic expressions for the voltage transfer characteristic.

6.2.1 Voltage Regime One: n-MOS Cutoff and p-MOS Linear

In voltage regime one, the n-MOS is cutoff, whereas the p-MOS is linear. There is negligible current flowing the p-MOS device and zero voltage drop across it, so that the output voltage is equal to V_{DD}:

$$V_{OH} = V_{DD} . \tag{6.1}$$

Figure 6.3 shows the load curve analysis for a CMOS inverter operating in regime one, with V_{DD} = 2.5 V, V_{TN} = 0.5 V, V_{TP} = −0.5 V, and V_{IN} = 0. I_{DD} = 0 for the n-MOS transistor, so its characteristic is coincident with the V_{OUT} axis.

6.2.2 Voltage Regime Two: n-MOS Saturated and p-MOS Linear

If the input voltage is increased slightly beyond the threshold voltage of the n-MOS device, it will be saturated but the p-MOSFET will remain linear. The condition for saturation operation of M_{NO} is

$$(V_{IN} \ge V_{TN}) \text{ and } (V_{IN} - V_{TN} \le V_{OUT}) . \tag{6.2}$$

If these conditions are satisfied, then the supply current is equal to the saturated drain current for the n-MOS transistor:

$$I_{DD} = \frac{K_N (V_{IN} - V_{TN})^2}{2} . \tag{6.3}$$

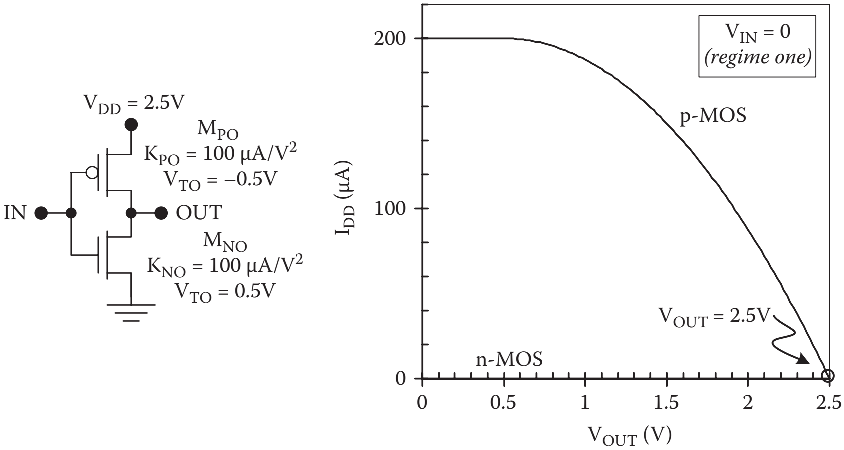


FIGURE 6.3
Load curve analysis for a symmetric CMOS inverter operating in voltage regime one with $V_{IN} = 0$. $V_{DD} = 2.5$ V, $K_{PO} = 100\ \mu A/V^2$, $V_{TP} = -0.5$ V, $K_{NO} = 100\ \mu A/V^2$, and $V_{TN} = 0.5$ V.

The output voltage can be determined using the drain-to-source voltage for the linear p-MOS device:

$$V_{OUT} = V_{DD} + V_{DSPO}\,, \tag{6.4}$$

where V_{DSPO} is negative and given by

$$\begin{aligned} V_{DSPO} &= (V_{GSPO} - V_{TP}) + \sqrt{(V_{GSPO} - V_{TP})^2 - \frac{2I_{DD}}{K_P}} \\ &= (V_{IN} - V_{DD} - V_{TP}) + \sqrt{(V_{IN} - V_{DD} - V_{TP})^2 - \frac{K_N}{K_P}(V_{IN} - V_{TN})^2}\,. \end{aligned} \tag{6.5}$$

Therefore,

$$V_{OUT} = (V_{IN} - V_{TP}) + \sqrt{(V_{IN} - V_{DD} - V_{TP})^2 - \frac{K_N}{K_P}(V_{IN} - V_{TN})^2}\,. \tag{6.6}$$

Because V_{OUT} is not known a priori, the voltage Conditions 6.2 must be checked for consistency after the calculation of V_{OUT} using Equations 6.6.

Figure 6.4 illustrates the load curve analysis for one particular operating point in regime two, for a symmetric CMOS inverter with $V_{DD} = 2.5$ V, $V_{TN} = 0.5$ V, $V_{TP} = -0.5$ V, and $V_{IN} = 1.2$. The solution is found with $V_{OUT} = 2.1$ V and $I_{DD} \sim 24\ \mu A$. From this graphical analysis, it is clear that the n-MOS transistor is saturated, whereas the p-MOSFET is linear at the point of intersection.

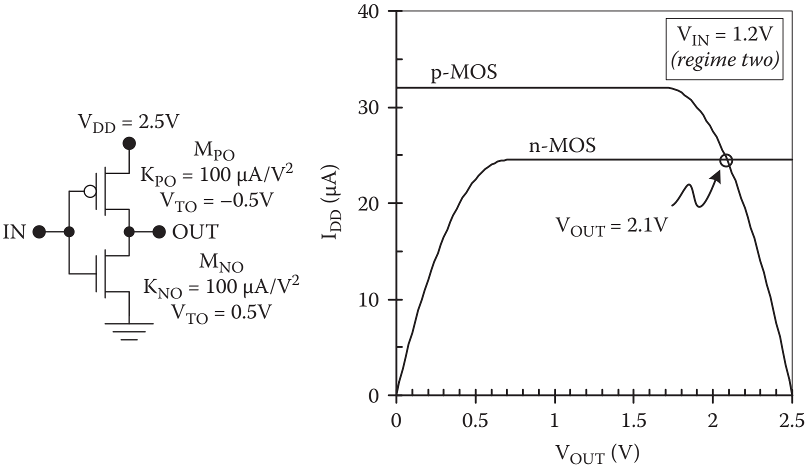


FIGURE 6.4
Load curve analysis for a symmetric CMOS inverter operating in voltage regime two with V_{IN} = 1.2V. V_{DD} = 2.5 V, K_{PO} = 100 µA/V^2, V_{TP} = –0.5V, K_{NO} = 100 µA/V^2, and V_{TN} = 0.5 V.

6.2.3 Voltage Regime Three: Both MOSFETs Saturated

If V_{IN} is further increased, the p-MOSFET will become saturated (voltage regime three). Therefore, both MOSFETs are saturated if

$$\left(V_{IN} - V_{TN} \le V_{OUT}\right) \text{ and } \left(V_{IN} - V_{TP} \ge V_{OUT}\right), \tag{6.7}$$

or, in a more practical form,

$$V_{TP} \le \left(V_{IN} - V_{OUT}\right) \le V_{TN}\ . \tag{6.8}$$

With both MOSFETs saturated, the exact voltage transfer characteristic cannot be calculated without knowledge of the channel length modulation parameters λ_N and λ_P. Equating the saturated drain currents yields

$$\left(K_N/2\right)\left(V_{IN} - V_{TN}\right)^2\left(1 + \lambda_N V_{OUT}\right) = \left(K_P/2\right)\left(V_{IN} - V_{DD} - V_{TP}\right)^2 \left(1 + \lambda_p\left(V_{DD} - V_{OUT}\right)\right) \tag{6.9}$$

and

$$V_{OUT} = \frac{K_P\left(V_{IN} - V_{DD} - V_{TP}\right)^2\left(1 + \lambda_P V_{DD}\right) - K_N\left(V_{IN} - V_{TN}\right)^2}{K_N\lambda_N\left(V_{IN} - V_{TN}\right)^2 + K_P\lambda_P\left(V_{IN} - V_{DD} - V_{TP}\right)^2}\ . \tag{6.10}$$

When the channel length modulation parameters are not known, good accuracy may be obtained by interpolating between the adjacent regions of the voltage transfer characteristic.

Figure 6.5 shows the load curve analysis for the case in which both channel length modulation parameters have been assume to be zero ($\lambda_N = \lambda_P = \lambda = 0$) and $V_{IN} = V_{DD}/2$. As in the previous examples, $V_{DD} = 2.5$ V, $V_{TN} = 0.5$ V, and $V_{TP} = -0.5$ V. With the assumption of $\lambda = 0$, the result is ambiguous, because there is a range of V_{OUT} for which the two drain currents are equal: $(V_{DD}/2 - V_{TN}) \leq V_{OUT} \leq (V_{DD}/2 - V_{TP})$. To estimate the solution, we might take the center of this range ($V_{OUT} = V_{DD}/2$), but if λ is small, the actual solution will be sensitive to variations in the other transistor parameters.

Figure 6.6 shows a more realistic load curve analysis for regime three using nonzero channel length modulation parameters ($\lambda_N = \lambda_P = \lambda = 0.1$ V^{-1}) but with $V_{IN} = V_{DD}/2$ as before. The result is no longer ambiguous; instead, the n-MOS and p-MOS characteristics intersect at a single operating point with $V_{OUT} = V_{DD}/2$ and I_{DD}= ~32 μA.

6.2.4 Voltage Regime Four: n-MOS Linear and p-MOS 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 voltage conditions in this regime are

$$(V_{IN} - V_{TN} \geq V_{OUT}) \text{ and } (V_{IN} \leq V_{DD} + V_{TP}). \tag{6.11}$$

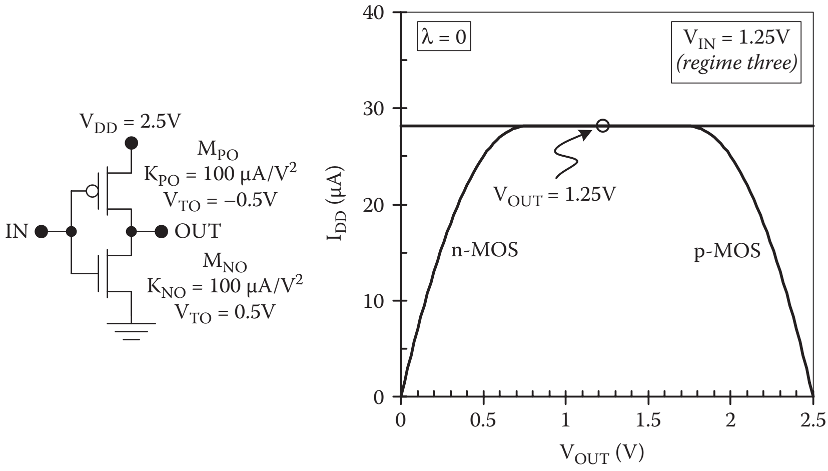


FIGURE 6.5

Load curve analysis for a symmetric CMOS inverter operating in voltage regime three with V_{IN} = 1.25V. V_{DD} = 2.5 V, K_{PO} = 100 μA/V^2, V_{TP} = −0.5 V, K_{NO} = 100 μA/V^2, and V_{TN} = 0.5V. Here, it was assumed that both channel length modulation parameters are equal to zero. ($\lambda_N = \lambda_P = \lambda = 0$.)

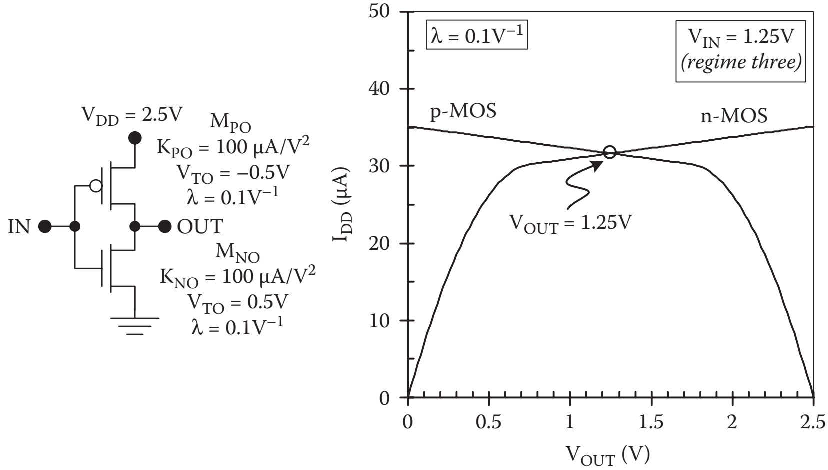


FIGURE 6.6
Load curve analysis for a symmetric CMOS inverter operating in voltage regime three with $V_{IN} = 1.25V$. $V_{DD} = 2.5$ V, $K_{PO} = 100$ μA/V^2, $V_{TP} = -0.5$ V, $K_{NO} = 100$ μA/V^2, and $V_{TN} = 0.5$ V. Here, it was assumed that $\lambda_N = \lambda_P = \lambda = 0.1$ V^{-1}.

Under these conditions,

$$I_{DD} = \frac{K_P\left(V_{IN} - V_{DD} - V_{TP}\right)^2}{2} \tag{6.12}$$

and

$$V_{OUT} = (V_{IN} - V_{TN}) - \sqrt{\left(V_{IN} - V_{TN}\right)^2 - \frac{K_P}{K_N}\left(V_{IN} - V_{DD} - V_{TP}\right)^2}\ . \tag{6.13}$$

Here, as in regime two, it is necessary to check the voltage Condition 6.10 after the determination of V_{OUT}.

The load curve analysis for an operating point from regime four is illustrated in Figure 6.7, with $V_{IN} = 1.3$ V and $V_{DD} = 2.5$ V. It is clear that the solution point ($V_{OUT} = 0.4$ V, $I_{DD} = 24$ μA) occurs on the flat (saturated) portion of the p-MOS characteristic but the sloping (linear) part of the n-MOS curve.

6.2.5 Voltage Regime Five: n-MOS Linear and p-MOS Cutoff

If the input voltage is sufficiently close to V_{DD}, then the p-MOS device will cut off. This occurs if

$$\left(V_{IN} \geq V_{DD} + V_{TP}\right) \tag{6.14}$$

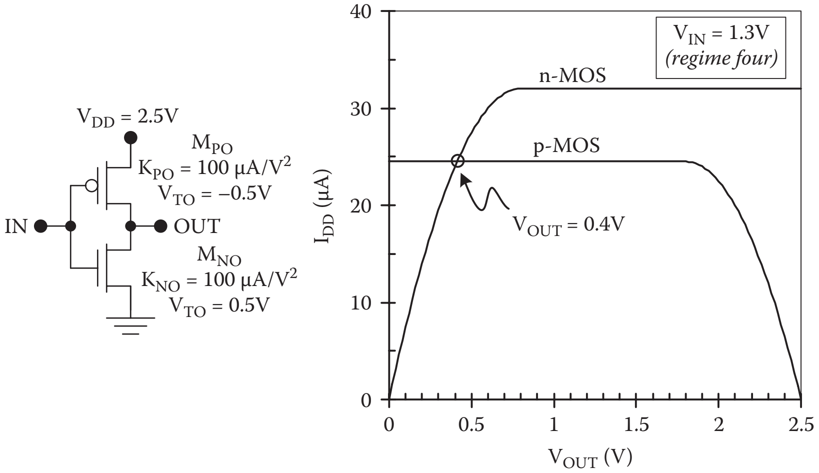


FIGURE 6.7
Load curve analysis for a symmetric CMOS inverter operating in voltage regime four with $V_{IN} = 1.3$ V. $V_{DD} = 2.5$ V, $K_{PO} = 100\ \mu A/V^2$, $V_{TP} = -0.5$ V, $K_{NO} = 100\ \mu A/V^2$, and $V_{TN} = 0.5$ V.

and results in zero voltage across the linear n-MOS transistor, so that

$$V_{OL} = 0\ . \tag{6.15}$$

CMOS circuits therefore provide *rail-to-rail voltage swing*, that is, the logic swing is equal to the supply voltage: $V_{OH} - V_{OL} = V_{DD}$.

Figure 6.8 displays the load curve analysis for one point within voltage regime five, with $V_{IN} = 2.0$ V and $V_{DD} = 2.5$ V. The p-MOSFET is cutoff so its characteristic is coincident with the V_{OUT} axis, and the solution corresponds to $V_{OUT} = 0$, $I_{DD} = 0$.

Example 6.1 CMOS Voltage Transfer Characteristic

Calculate the voltage transfer characteristic for a symmetric CMOS inverter of Figure 6.9 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.

$$V_{OUT} = \begin{cases} 2.5V; & \text{(regime 1;} \quad V_{IN} \le 0.5V) \\ V_{IN} + 0.5V + \sqrt{(V_{IN} - 2.0V)^2 - (V_{IN} - 0.5V)^2}; & \text{(regime 2;} \quad 0.5V \le V_{IN} \le 1.25V) \\ 1.25V; & \text{(regime 3;} \quad V_{IN} \approx 1.25V) \\ V_{IN} - 0.5V - \sqrt{(V_{IN} - 0.5V)^2 - (V_{IN} - 2.0V)^2}; & \text{(regime 4;} \quad 1.25V \le V_{IN} \le 2.0V) \\ 0; & \text{(regime 5;} \quad V_{IN} \ge 2.0V). \end{cases}$$

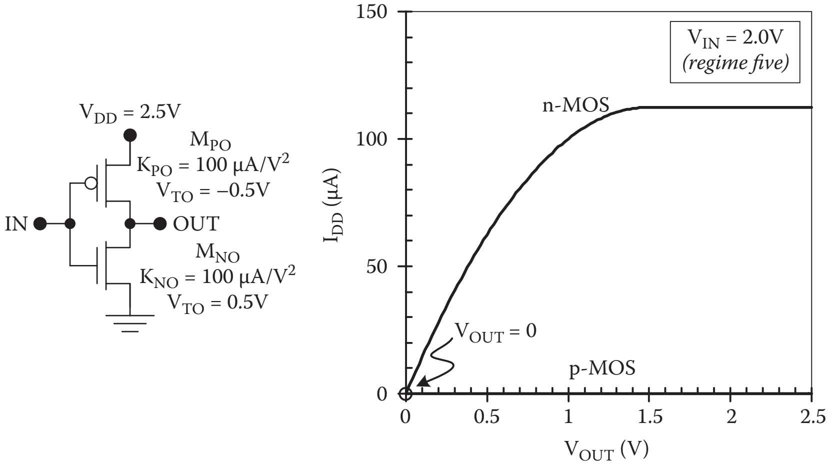


FIGURE 6.8
Load curve analysis for a symmetric CMOS inverter operating in voltage regime four with $V_{IN} = 1.3$ V. $V_{DD} = 2.5$ V, $K_{PO} = 100\ \mu A/V^2$, $V_{TP} = -0.5$ V, $K_{NO} = 100\ \mu A/V^2$, and $V_{TN} = 0.5$ V.

Figure 6.10 shows the calculated voltage transfer characteristic with the five voltage regimes labeled.

6.3 Load Surface Analysis

The load surface analysis described in Chapter 5 may also be applied to a CMOS inverter. For this analysis, the drain currents in the n-MOS and p-MOS transistors are given by

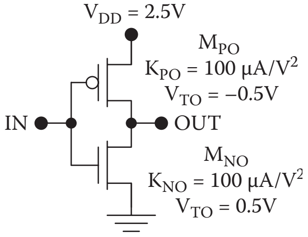


FIGURE 6.9
Example CMOS inverter for the calculation of the VTC.

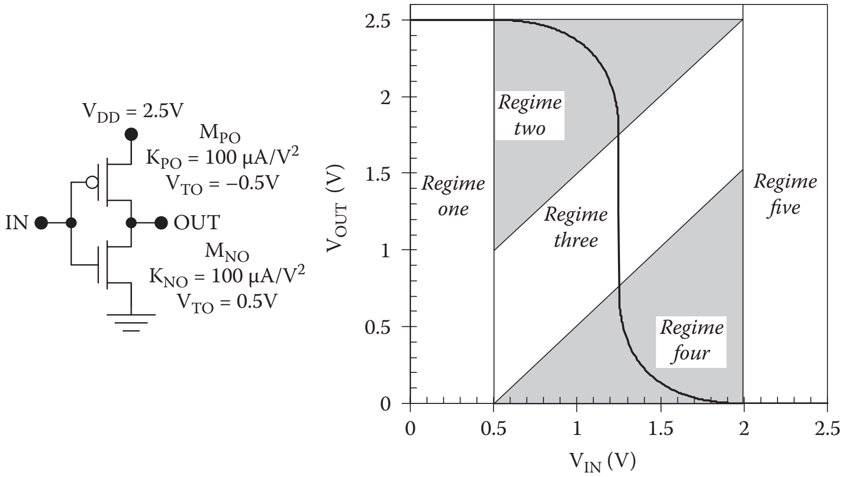


FIGURE 6.10
Calculated VTC for a symmetric CMOS inverter with K_{PO} = 100 µA/V^2, V_{TP} = −0.5 V, K_{NO} = 100 µA/V^2, V_{TN} = 0.5 V, and V_{DD} = 2.5 V.

$$I_{DN} = \begin{cases} 0; \\ K_{NO}\left(V_{IN} - V_{TN}\right)^2 / 2; \\ K_{NO}\left[\left(V_{IN} - V_{TN}\right)V_{OUT} - V_{OUT}^2 / 2\right]; \end{cases}$$

$$\begin{matrix} 0 \le V_{IN} \le V_{TN} & \text{(cutoff)} \\ V_{TN} \le V_{IN} \le \left(V_{OUT} + V_{TN}\right) & \text{(saturation)} \\ \left(V_{OUT} + V_{TN}\right) \le V_{IN} & \text{(linear)} \end{matrix} \tag{6.16}$$

and

$$I_{DP} = \begin{cases} K_{PO}\left[\left(V_{IN} - V_{DD} - V_{TP}\right)\left(V_{OUT} - V_{DD}\right) - \left(V_{OUT} - V_{DD}\right)^2 / 2\right] \\ K_{PO}\left(V_{IN} - V_{DD} - V_{TP}\right)^2 / 2; \\ 0; \end{cases}$$

$$\begin{matrix} V_{IN} \le V_{OUT} + V_{TP} & \text{(linear)} \\ \left(V_{OUT} + V_{TP}\right) \le V_{IN} \le \left(V_{DD} + V_{TP}\right) & \text{(saturation)} \\ \left(V_{DD} + V_{TP}\right) \le V_{IN} & \text{(cutoff).} \end{matrix} \tag{6.17}$$

Figure 6.11 shows the load surface analysis for a symmetric CMOS inverter with K_{PO} = 100 µA/V^2, V_{TP} = −0.5 V, K_{NO} = 100 µA/V^2, V_{TN} = 0.5 V, and V_{DD} = 2.5 V.

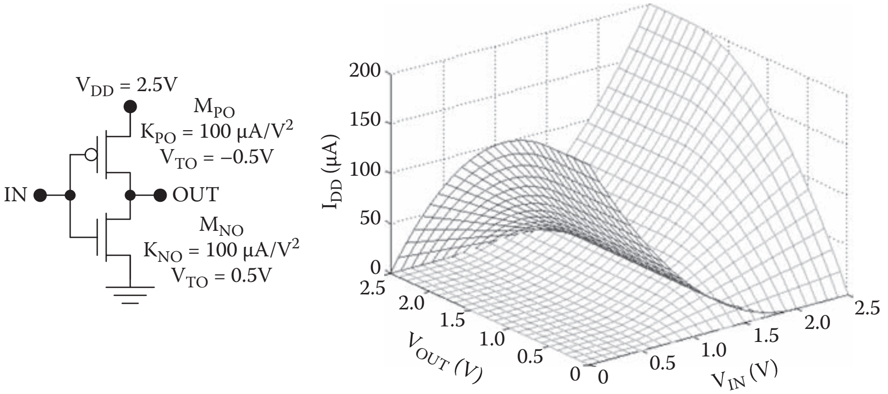


FIGURE 6.11
Load surface analysis for a symmetric CMOS inverter with K_{PO} = 100 µA/V², V_{TP} = –0.5V, K_{NO} = 100 µA/V², V_{TN} = 0.5 V, and V_{DD} = 2.5 V.

The voltage transfer characteristic, determined from the projection of the load surface solution into the voltage plane, is shown in Figure 6.12 with the modes of operation indicated.

6.4 Critical Voltages

It has already been shown that CMOS circuits exhibit rail-to-rail swing under static conditions, with $V_{OL} = 0$ and $V_{OH} = V_{DD}$. In this section, we will consider the other critical voltages V_{IL}, V_M, and V_{IH}. The input low-voltage V_{IL} is the maximum input voltage that will be interpreted as logic zero, and by definition, it is the input voltage for which $dV_{OUT}/dV_{IN} = -1$. This operating point exists in voltage regime two with the n-MOS transistor saturated as indicated in Figure 6.13. Similarly, the input high-voltage V_{IH} is the minimum input voltage that will be interpreted as logic one, and it occurs in voltage regime four with the p-MOS transistor saturated. The switching threshold, or midpoint voltage, V_M is the value of the input voltage for which the input and output are equal and occurs in voltage regime three with both transistors saturated.

6.4.1 Input Low-Voltage V_{IL}

The input low-voltage occurs in regime two with the n-MOS transistor saturated but the p-MOS transistor linear. The drain currents are given by

$$I_{DN} = \frac{K_N}{2}\left(V_{IN} - V_{TN}\right)^2 \tag{6.18}$$

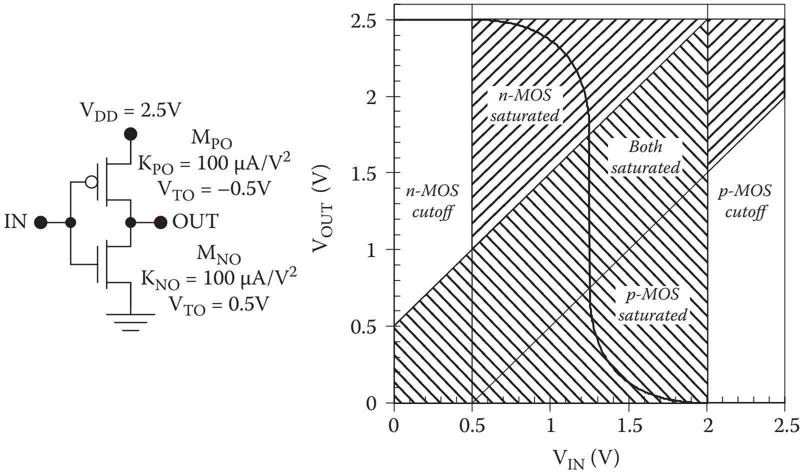


FIGURE 6.12
Voltage transfer characteristic for a symmetric CMOS inverter with $K_{PO} = 100\ \mu A/V^2$, $V_{TP} = -0.5$ V, $K_{NO} = 100\ \mu A/V^2$, $V_{TN} = 0.5$ V, and $V_{DD} = 2.5$ V with the device modes of operation indicated.

and

$$I_{DP} = K_P\left[(V_{IN} - V_{DD} - V_{TP})(V_{OUT} - V_{DD}) - \frac{(V_{OUT} - V_{DD})^2}{2}\right]. \tag{6.19}$$

Equating the drain currents, $I_{DN} = I_{DP}$, $dI_{DO} = dI_{DL}$, and

$$\frac{\partial I_{DN}}{\partial V_{IN}} dV_{IN} = \frac{\partial I_{DP}}{\partial V_{IN}} dV_{IN} + \frac{\partial I_{DP}}{\partial V_{OUT}} dV_{OUT}. \tag{6.20}$$

The slope of the transfer characteristic can therefore be determined using the partial derivatives:

$$\frac{dV_{OUT}}{dV_{IN}} = \frac{\dfrac{\partial I_{DN}}{\partial V_{IN}} - \dfrac{\partial I_{DP}}{\partial V_{IN}}}{\dfrac{\partial I_{DP}}{\partial V_{OUT}}} = \frac{K_N(V_{IN} - V_{TN}) - K_P(V_{OUT} - V_{DD})}{K_P(V_{IN} - V_{OUT} - V_{TP})}. \tag{6.21}$$

By definition, $dV_{OUT}/dV_{IN} = -1$ at the input low voltage; therefore,

$$V_{IL} = \frac{2V_{OUT} - V_{DD} + V_{TP} + (K_N / K_P)V_{TN}}{1 + K_N / K_P}. \tag{6.22}$$

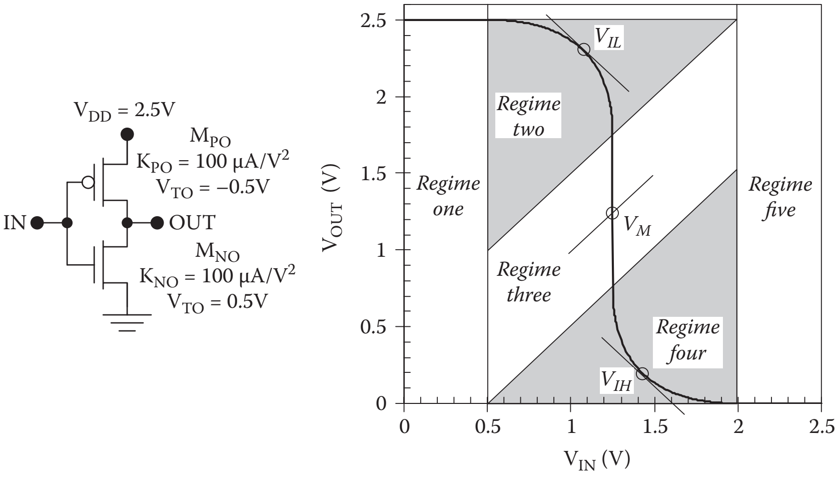


FIGURE 6.13
Critical input voltages V_{IL}, V_M, and V_{IH} for a symmetric CMOS inverter.

The application of this relationship is complicated by the fact that V_{OUT} is a function of V_{IL}. Making use of the regime two equation,

$$V_{OUT} = (V_{IL} - V_{TP}) + \sqrt{(V_{IL} - V_{DD} - V_{TP})^2 - \frac{K_N}{K_P}(V_{IL} - V_{TN})^2}\,. \tag{6.23}$$

For approximate hand calculations, we can assume $V_{OUT}(V_{IL}) \approx 0.9V_{DD}$. Then

$$V_{IL} \approx \frac{4V_{DD}/5 + V_{TP} + (K_N/K_P)V_{TN}}{1 + K_N/K_P}\,. \tag{6.24}$$

6.4.2 Switching Threshold V_M

The most important critical voltage for CMOS design is the switching threshold, or midpoint voltage, for which $V_{OUT} = V_{IN} = V_M$. This point on the voltage transfer characteristic occurs in regime three and may be determined approximately by equating the saturated drain currents with the assumption that $\lambda_N = \lambda_P = 0$. Then,

$$(K_N/2)(V_M - V_{TN})^2 = (K_P/2)(V_M - V_{DD} - V_{TP})^2 \tag{6.25}$$

and

$$V_M = \frac{V_{TN} + (V_{DD} + V_{TP})\sqrt{K_P/K_N}}{1 + \sqrt{K_P/K_N}}\,. \tag{6.26}$$

This may also be written in terms of the transconductance ratio $K_R = K_N / K_P$:

$$V_M = \frac{V_{TN} + (V_{DD} + V_{TP})\sqrt{1/K_R}}{1 + \sqrt{1/K_R}}. \tag{6.27}$$

Therefore, the switching threshold varies with the ratio of the transistor widths even with fixed threshold voltages and a fixed supply voltage as shown in Figure 6.14. Although this result was obtained by assuming $\lambda_N = \lambda_P = 0$, the switching threshold is only weakly affected by the channel length modulation parameter.

When designing a CMOS inverter to achieve a desired switching threshold, rearrangement of Equation 6.27 gives the design equation

$$K_R = \left(\frac{V_{DD} + V_{TP} - V_M}{V_M - V_{TN}}\right)^2. \tag{6.28}$$

6.4.3 Input High-Voltage V_{IH}

The input high voltage may be found in similar manner to V_{IL}; however, this point occurs in voltage regime four so we start with the assumptions that the n-MOSFET is linear and the p-MOSFET is saturated. The drain currents are

$$I_{DN} = K_N\left[(V_{IN} - V_{TN})V_{OUT} - V_{OUT}^2 / 2\right] \tag{6.29}$$

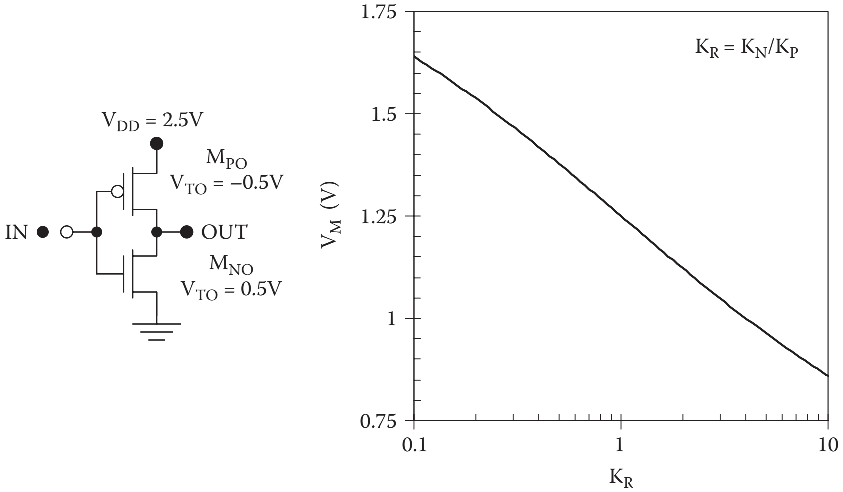


FIGURE 6.14
Switching threshold V_M as a function of the transconductance ratio K_R for a CMOS inverter with V_{DD} = 2.5 V, V_{TN} = 0.5 V, and V_{TP} = −0.5 V.

and

$$I_{DP} = K_P\left(V_{IN} - V_{DD} - V_{TP}\right)^2 / 2\,. \tag{6.30}$$

As before, $I_{DN} = I_{DP}$ and $dI_{DO} = dI_{DL}$ so that

$$\frac{\partial I_{DN}}{\partial V_{IN}} dV_{IN} + \frac{\partial I_{DN}}{\partial V_{OUT}} dV_{OUT} = \frac{\partial I_{DP}}{\partial V_{IN}} dV_{IN}\,. \tag{6.31}$$

At the input high voltage,

$$\frac{dV_{OUT}}{dV_{IN}} = \frac{\dfrac{\partial I_{DP}}{\partial V_{IN}} - \dfrac{\partial I_{DN}}{\partial V_{IN}}}{\dfrac{\partial I_{DN}}{\partial V_{OUT}}} = \frac{K_P\left(V_{IN} - V_{DD} - V_{TP}\right) - K_N V_{OUT}}{K_N\left(V_{IN} - V_{TN} - V_{OUT}\right)} = -1\,. \tag{6.32}$$

Solving for V$_{IN}$, we obtain

$$V_{IH} = \frac{2V_{OUT} + V_{TN} + \left(K_P / K_N\right)\left(V_{DD} + V_{TP}\right)}{1 + K_P / K_N}\,. \tag{6.33}$$

If we make the approximation that $V_{OUT}\left(V_{IH}\right) \approx V_{DD} / 10$, then

$$V_{IH} \approx \frac{V_{OUT} / 5 + V_{TN} + \left(K_P / K_N\right)\left(V_{DD} + V_{TP}\right)}{1 + K_P / K_N}\,. \tag{6.34}$$

6.5 Crossover (Short-Circuit) Current

In either normal output state for the CMOS logic circuit, one of the two MOS transistors is cutoff so the supply current is approximately zero. However, significant crossover current (also known as short-circuit current) may flow during switching transitions attributable to the simultaneous conduction of the pull-down and pull-up circuits. In the CMOS inverter, the crossover current characteristic may be calculated piecewise for four regimes of operation. For this purpose, we will consider the unloaded CMOS inverter of Figure 6.15 with the approximation that both channel length modulation parameters are zero and assuming that the subthreshold currents may be neglected.

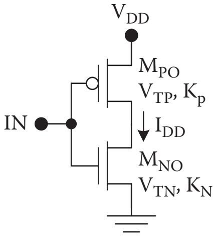


FIGURE 6.15
CMOS inverter for estimation of the crossover current.

6.5.1 Current Regime One: n-MOS Cutoff

If the input voltage is less than V_{TN}, the n-MOSFET will be cutoff. Because of this, no current will flow although the p-MOS transistor is operating in the linear mode:

$$I_{DD} = 0; \qquad (V_{IN} \le V_{TN}). \tag{6.35}$$

6.5.2 Current Regime Two: n-MOS Saturated

For $V_{TN} \le V_{IN} \le V_M$, the n-MOS transistor is saturated; it therefore acts as a voltage-controlled current source and sets the level of the crossover current:

$$I_{DD} = \frac{K(V_{IN} - V_{TN})^2}{2}; \qquad (V_T \le V_{IN} \le V_M). \tag{6.36}$$

6.5.3 Current Regime Three: p-MOS Saturated

For the input voltage in the range $V_M \le V_{IN} \le (V_{DD} + V_{TP})$, the p-MOS transistor is saturated and limits the crossover current:

$$I_{DD} = \frac{K(V_{IN} - V_{DD} - V_{TP})^2}{2}; \qquad (V_M \le V_{IN} \le (V_{DD} - V_T)). \tag{6.37}$$

6.5.4 Current Regime Four: p-MOS Cutoff

If the input voltage is more positive than $V_{DD} + V_{TP}$, the p-MOS device is cutoff so zero current flows:

$$I_{DD} = 0; \qquad (V_{IN} \ge V_{DD} + V_{TP}). \tag{6.38}$$

6.5.5 Unified Expression for the Crossover Current

Across current regimes two and three, the crossover current may be found as the minimum of Equations 6.35 and 6.36, without the need for calculating the switching threshold V_M. This yields a unified expression for the crossover current:

$$I_{DD} = \begin{cases} 0; & V_{IN} \le V_{TN} \\ \min\left\{K_N\left(V_{IN} - V_{TN}\right)^2/2,\ K_P\left(V_{IN} - V_{DD} + V_{TP}\right)^2/2\right\} & V_{TN} \le V_{IN} \le \left(V_{DD} + V_{TP}\right) \\ 0 & \left(V_{DD} + V_{TP}\right) \le V_{IN} \end{cases} \quad (6.39)$$

Example 6.2 Crossover Current in a Symmetric Inverter

Calculate the crossover current for the symmetric CMOS inverter of Figure 6.16 with VDD = 2.5 V, VTN = |VTP| = 0.5 V, and KN = KP = 100 µA/V2.

Solution: The crossover current is given by

$$I_{DD} = \begin{cases} 0; & (\text{regime 1};\quad V_{IN} \le 0.5V) \\ 100\mu\text{A/V}^2\left(V_{IN} - 0.5V\right)^2/2; & (\text{regime 2};\quad 0.5V \le V_{IN} \le 1.25V) \\ 100\mu\text{A/V}^2\left(V_{IN} - 2.0V\right)^2/2; & (\text{regime 3};\quad 1.25V \le V_{IN} \le 2.0V) \\ 0; & (\text{regime 4};\quad 2.0V \le V_{IN}) \end{cases}.$$

The symmetric characteristic is shown in Figure 6.17 with the current regimes indicated. The peak crossover current is $I_{peak} = 100\mu A/V^2\left(1.25V - 0.5V\right)^2/2 = 28\mu A$.

6.5.6 Effect of Threshold Voltages

From Equation 6.39, it can be seen that crossover current only flows for the input voltage range $V_{TN} \le V_{IN} \le \left(V_{DD} + V_{TP}\right)$. This range can be restricted

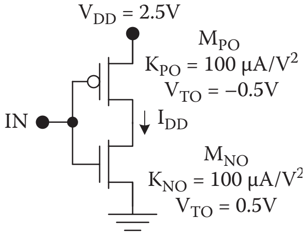


FIGURE 6.16
Example CMOS inverter for the calculation of the crossover current.

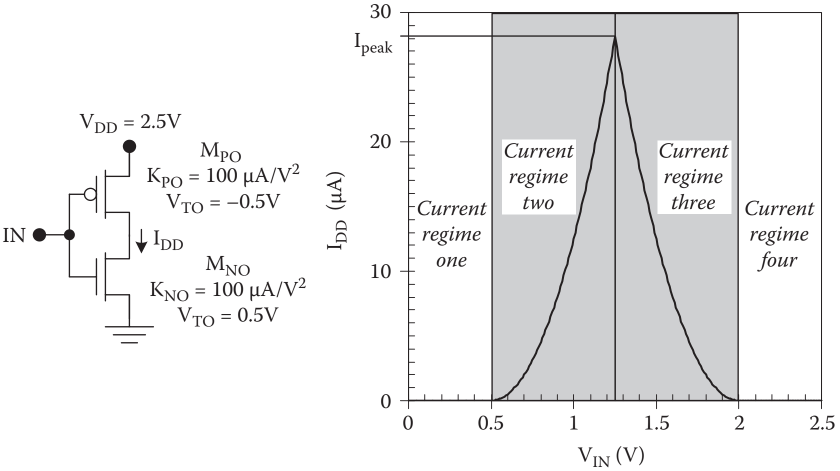


FIGURE 6.17
Calculated crossover current as a function of input voltage for a symmetric CMOS inverter with $K_P = 100\ \mu A/V^2$, $V_{TP} = -0.5$ V, $K_N = 100\ \mu A/V^2$, $V_{TN} = 0.5$ V, and $V_{DD} = 2.5$ V.

by increasing the absolute values of the threshold voltages as shown in Figure 6.17. In fact, the crossover current may be eliminated entirely if $\left(V_{TN} + \left|V_{TP}\right|\right) \geq V_{DD}$. However, such an approach would involve a tradeoff in switching speed because both transistors would have compromised current drive capability.

Example 6.3 Effect of Threshold Voltage on Crossover Current

Calculate the crossover current for symmetric CMOS inverters with $K_N = K_P = 100\ \mu A/V^2$, $V_{DD} = 2.5$ V, and $V_{TN} = |V_{TP}| = V_T$, for the cases of $V_T = 0.5$ V, 0.75 V, and 1.00 V.

Solution: The crossover current is given by

$$I_{DD} = \begin{cases} 0; & (\text{regime 1;} \quad V_{IN} \leq V_T) \\ 100\mu\text{A/V}^2 \left(V_{IN} - V_T\right)^2 / 2; & (\text{regime 2;} \quad V_T \leq V_{IN} \leq 1.25V) \\ 100\mu\text{A/V}^2 \left(V_{IN} - 2.5V + V_T\right)^2 / 2; & (\text{regime 3;} \quad 1.25V \leq V_{IN} \leq V_T) \\ 0; & (\text{regime 4;} \quad (2.5V - V_T) \leq V_{IN}) \end{cases}.$$

The characteristics are shown in Figure 6.18, and it can be seen that the peak crossover current decreases with increasing V_T.

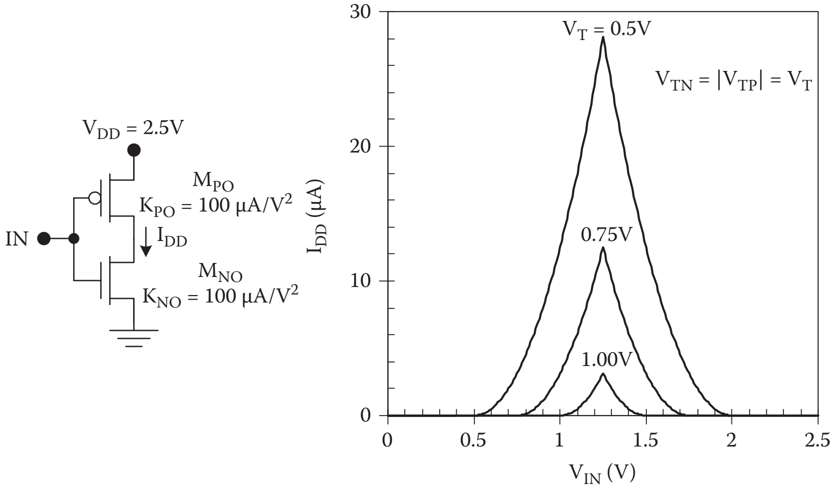


FIGURE 6.18
Crossover current I_{DD} as a function of the input voltage, with the absolute value of the threshold voltages as a parameter. $V_{DD} = 2.5$ V, $K_{NO} = K_{PO} = 100\ \mu A/V^2$, and $\lambda_N = \lambda_P = 0$.

Example 6.4 Crossover Current in Symmetric and symmetric CMOS Inverters

Calculate the short circuit current versus the input voltage for the symmetric and minimum-size inverter circuits shown in Figure 6.19. $V_{DD} = 2.5$ V, $V_{TN} = 0.5$ V, $V_{TP} = -0.5$ V, and $t_{OX} = 9$ nm.

Solution: The process transconductance values for the p-MOS and n-MOS transistors are

$$k'_P = \frac{\mu_p \varepsilon_{OX}}{t_{OX}} = \frac{\left(230cm^2/Vs\right)(3.9)\left(8.85\times10^{-14}F/cm\right)}{9\times10^{-7}cm} = 88\mu A/V^2$$

and

$$k'_N = \frac{\mu_n \varepsilon_{OX}}{t_{OX}} = \frac{\left(580cm^2/Vs\right)(3.9)\left(8.85\times10^{-14}F/cm\right)}{9\times10^{-7}cm} = 220\mu A/V^2 \cdot$$

For the symmetric CMOS inverter,

$$K_P = k'_P\left(\frac{W_P}{L_P}\right) = 88\mu A/V^2\left(\frac{3.0\mu m}{0.6\mu m}\right) = 440\mu A/V^2$$

and

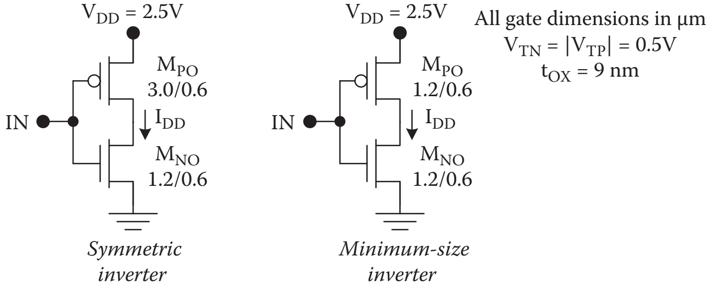


FIGURE 6.19
Example symmetric and minimum-size CMOS inverter circuits for the calculation of the crossover current.

$$K_N = k_N'\left(\frac{W_N}{L_N}\right) = 220\mu A/V^2\left(\frac{1.2\mu m}{0.6\mu m}\right) = 440\mu A/V^2.$$

The crossover current is given by

$$I_{DD} = \begin{cases} 0; & V_{IN} \le 0.5V \\ \min\left\{220\mu A/V^2\left(V_{IN}-0.5V\right)^2, 220\mu A/V^2\left(V_{IN}-2.0V\right)^2\right\} & 0.5V \le V_{IN} \le 2.0V \\ 0 & 2.0V \le V_{IN} \end{cases}$$

For the minimum-size CMOS inverter,

$$K_P = k_P'\left(\frac{W_P}{L_P}\right) = 88\mu A/V^2\left(\frac{1.2\mu m}{0.6\mu m}\right) = 176\mu A/V^2$$

and

$$K_N = k_N'\left(\frac{W_N}{L_N}\right) = 220\mu A/V^2\left(\frac{1.2\mu m}{0.6\mu m}\right) = 440\mu A/V^2.$$

The crossover current in the minimum-size inverter is given by

$$I_{DD} = \begin{cases} 0; & V_{IN} \le 0.5V \\ \min\left\{220\mu A/V^2\left(V_{IN}-0.5V\right)^2, 88\mu A/V^2\left(V_{IN}-2.0V\right)^2\right\} & 0.5V \le V_{IN} \le 2.0V. \\ 0 & 2.0V \le V_{IN} \end{cases}$$

As shown in Figure 6.20, the symmetric inverter exhibits peak crossover current at $V_{DD}/2$. The minimum-size inverter has compromised current drive in the p-MOS transistor and exhibits 40% lower peak crossover current at an input voltage of ~1.1 V.

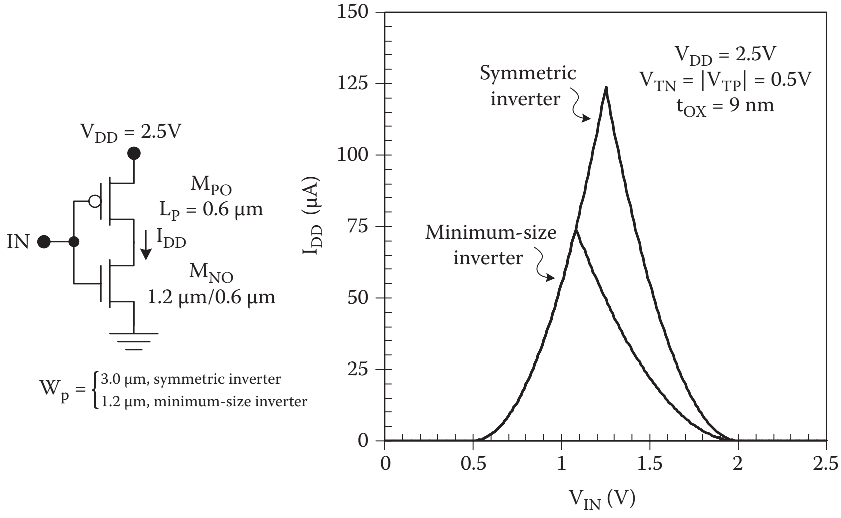


FIGURE 6.20
Calculated crossover current as a function of the input voltage for the symmetric and minimum-size inverter circuits of Figure 6.19.

6.6 Propagation Delays

To estimate the propagation delays for CMOS circuits, we will consider an inverter with a lumped capacitive load and abrupt voltage transitions at the input. In Section 6.6.6, we will consider the effect of the input rise/fall time, and, in Sections 6.14 and 6.15, we will briefly describe how the circuit design (NAND, NOR) affects the propagation delays. The effect of a distributed load will be considered in Chapter 7.

6.6.1 High-to-Low Propagation Delay t_{PHL}

Consider a symmetric CMOS inverter with a lumped capacitive load as shown in Figure 6.21 and suppose that the input voltage makes an abrupt transition from zero to V_{DD} at $t = 0$.

The n-MOS transistor becomes saturated at $t = 0^+$ and remains in this mode of operation until V_{OUT} drops to $V_{DD} - V_{TN}$. During this time interval, a constant current flows in M_{NO}:

$$I_{DN} = \frac{K_N \left(V_{DD} - V_{TN}\right)^2}{2}. \tag{6.40}$$

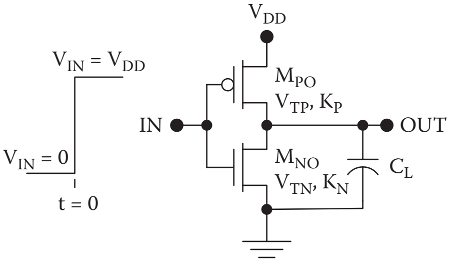


FIGURE 6.21
CMOS inverter with a lumped capacitive load and an abrupt low-to-high voltage transition at the input for the estimation of t_{PHL}.

The n-MOSFET becomes linear when $V_{OUT} = (V_{DD} - V_{TN})$ at $t = t_{PHL1}$:

$$t_{PHL1} = \frac{2V_{TN}C_L}{K_N\left(V_{DD} - V_{TN}\right)^2} . \tag{6.41}$$

After V_{OUT} drops below $V_{DD} - V_T$, the n-MOS operates in the linear mode until the end of the high-to-low propagation delay; by definition, this is the time at which $V_{OUT} = (V_{OH} + V_{OL})/2 = V_{DD}/2$. The length of this time interval for linear operation of the n-MOS transistor is t_{PHL2}, given by

$$\begin{aligned} t_{PHL2} &= -C_L \int_{V_{DD}-V_{TN}}^{V_{DD}/2} \frac{dV_{OUT}}{K_N\left[\left(V_{DD} - V_{TN}\right)V_{OUT} - V_{OUT}^2 / 2\right]} \\ &= -\frac{C_L}{K_N\left(V_{DD} - V_{TN}\right)}\left[ln\left(\frac{V_{OUT}}{V_{OUT} - 2\left(V_{DD} - V_{TN}\right)}\right)\right|_{V_{DD}-V_{TN}}^{V_{DD}/2} \\ &= \frac{C_L}{K_N\left(V_{DD} - V_{TN}\right)}\ln\left(\frac{3V_{DD} - 4V_{TN}}{V_{DD}}\right). \end{aligned} \tag{6.42}$$

The high-to-low propagation delay is given by the sum $t_{PHL1} + t_{PHL2}$, so that

$$t_{PHL} = \frac{C_L}{K_N\left(V_{DD} - V_{TN}\right)}\left[\frac{2V_{TN}}{\left(V_{DD} - V_{TN}\right)} + \ln\left(\frac{3V_{DD} - 4V_{TN}}{V_{DD}}\right)\right]. \tag{6.43}$$

Therefore, the high-to-low propagation delay is proportional to the load capacitance and inversely proportional to the n-MOS device transconductance parameter. Also, as a first-order approximation, the propagation delay varies inversely with $(V_{DD} - V_{TN})$.

6.6.2 Low-to-High Propagation Delay t_{PLH}

For the determination of t_{PLH}, we will assume that the input voltage makes an abrupt high-to-low transition as shown in Figure 6.22.

The p-MOS transistor becomes saturated at $t = 0^+$ and remains in this mode of operation until V_{OUT} increases to $|V_{TP}|$. The time interval t_{PLH1} for saturated operation of the p-MOS transistor is given byz

$$t_{PLH1} = -\frac{2V_{TP}C_L}{K_P\left(V_{DD}+V_{TP}\right)^2}\,. \tag{6.44}$$

The p-MOS transistor will operate in the linear mode for the rest of the propagation delay, during a time interval t_{PLH2} given by

$$\begin{aligned} t_{PLH2} &= C_L\int_{-V_{TP}}^{V_{DD}/2}\frac{dV_{OUT}}{K_P\left[\left(V_{DD}+V_{TP}\right)\left(V_{DD}-V_{OUT}\right)-\left(V_{DD}-V_{OUT}\right)^2/2\right]} \\ &= \frac{C_L}{K_P\left(V_{DD}+V_{TP}\right)}\left[ln\left(\frac{-V_{DD}-2V_{TP}-V_{OUT}}{V_{DD}-V_{OUT}}\right)\Bigg|_{-V_{TP}}^{V_{DD}/2}\right. \\ &= \frac{C_L}{K_P\left(V_{DD}+V_{TP}\right)}\ln\left(\frac{3V_{DD}+4V_{TP}}{V_{DD}}\right). \end{aligned} \tag{6.45}$$

The low-to-high propagation delay is the sum $t_{PLH1} + t_{PHL2}$, which is

$$t_{PLH} = \frac{C_L}{K_P\left(V_{DD}+V_{TP}\right)}\left[\frac{-2V_{TP}}{\left(V_{DD}+V_{TP}\right)}+\ln\left(\frac{3V_{DD}+4V_{TP}}{V_{DD}}\right)\right]. \tag{6.46}$$

Thus, t_{PLH} is proportional to the load capacitance C_L, is inversely proportional to K_P, and (as a first-order approximation) varies inversely with $(V_{DD} + V_{TP})$.

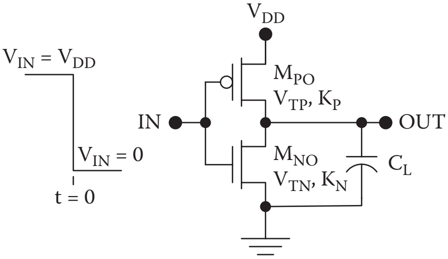


FIGURE 6.22
CMOS inverter with a lumped capacitive load and an abrupt low-to-high voltage transition at the input for the estimation of t_{PHL}.

6.6.3 Propagation Delay Design Equations

By rearranging Equations 6.43 and 6.46, we can develop equations that allow us to choose the aspect ratios and therefore gate widths of the transistors to achieve the necessary speed performance. If the p-MOS transistor must be sized to achieve a low-to-high propagation delay better than $t_{PLH,\max}$ with a load capacitance equal to C_L, then the design equation is

$$\frac{W_P}{L_P} \geq \frac{C_L}{\mu_p C_{OX} t_{PLH,\max}\left(V_{DD}+V_{TP}\right)}\left[\frac{-2V_{TP}}{\left(V_{DD}+V_{TP}\right)}+\ln\left(\frac{3V_{DD}+4V_{TP}}{V_{DD}}\right)\right] \quad (6.47)$$

or

$$\frac{W_P}{L_P} \geq \frac{C_L}{\mu_p C_{OX} t_{PLH,\max}}\Gamma_P, \quad (6.48)$$

where

$$\Gamma_P = \frac{1}{\left(V_{DD}+V_{TP}\right)}\left[\frac{-2V_{TP}}{\left(V_{DD}+V_{TP}\right)}+\ln\left(\frac{3V_{DD}+4V_{TP}}{V_{DD}}\right)\right]. \quad (6.49)$$

Similarly, if the n-MOS transistor is to be sized for a maximum high-to-low propagation delay $t_{PHL,\max}$ with a load capacitance equal to C_L, then the design equation is

$$\frac{W_N}{L_N} \geq \frac{C_L}{\mu_p C_{OX} t_{PHL,\max}\left(V_{DD}-V_{TN}\right)}\left[\frac{2V_{TN}}{\left(V_{DD}-V_{TN}\right)}+\ln\left(\frac{3V_{DD}-4V_{TN}}{V_{DD}}\right)\right] \quad (6.50)$$

or

$$\frac{W_N}{L_N} \geq \frac{C_L}{\mu_n C_{OX} t_{PHL,\max}}\Gamma_N, \quad (6.51)$$

where

$$\Gamma_N = \frac{1}{\left(V_{DD}-V_{TN}\right)}\left[\frac{2V_{TN}}{\left(V_{DD}-V_{TN}\right)}+\ln\left(\frac{3V_{DD}-4V_{TN}}{V_{DD}}\right)\right]. \quad (6.52)$$

6.6.4 Propagation Delays in the Symmetric Inverter

For the symmetric inverter, in which $K_P = K_N = K$ and $V_{TN} = |V_{TP}| = V_T$, the propagation delays are equal, and

$$t_{PLH} = t_{PHL} = t_P = \frac{C_L}{K\left(V_{DD}-V_T\right)}\left[\frac{2V_T}{\left(V_{DD}-V_T\right)}+\ln\left(\frac{3V_{DD}-4V_T}{V_{DD}}\right)\right]. \quad (6.53)$$

6.6.5 Approximate Expressions for the Propagation Delays

For a symmetric CMOS inverter, the propagation delays may be estimated (usually with better than 25% accuracy) using

$$t_P \approx \frac{1.6C_L}{K\left(V_{DD}-V_T\right)}. \tag{6.54}$$

This expression also correctly predicts $t_P \propto C_L$, $t_P \propto 1/K$, and $t_P \propto 1/(V_{DD}-V_T)$. For asymmetric circuits, the propagation delays may be estimated by

$$t_{PLH} \approx \frac{1.6C_L}{K_P\left(V_{DD}+V_{TP}\right)} \tag{6.55}$$

and

$$t_{PHL} \approx \frac{1.6C_L}{K_N\left(V_{DD}-V_{TN}\right)}. \tag{6.56}$$

Improved speed can be obtained by reduction of the load capacitance or by scaling up the K values; both of these may be accomplished by scaling down the channel lengths of the MOS transistors.

Example 6.5 Propagation Delays for the Symmetric CMOS Inverter

Estimate t_P (1 pF load) for the symmetric CMOS inverter depicted in Figure 6.23 with V_{DD} = 2.5 V, V_T = 0.5 V, and t_{OX} = 9 nm, using the detailed and approximate equations.

Solution: The process transconductance values for the p-MOS and n-MOS transistors are

$$k_P' = \frac{\mu_p \varepsilon_{OX}}{t_{OX}} = \frac{\left(230cm^2/Vs\right)(3.9)\left(8.85\times10^{-14}F/cm\right)}{9\times10^{-7}cm} = 88\mu A/V^2$$

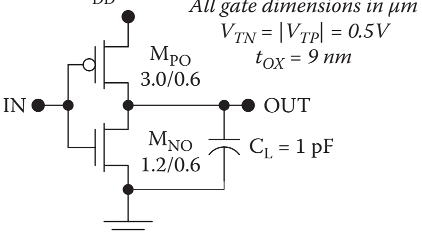


FIGURE 6.23
Example symmetric inverter for the calculation of the propagation delays.

and

$$k'_N = \frac{\mu_n \varepsilon_{OX}}{t_{OX}} = \frac{\left(580cm^2 / Vs\right)(3.9)\left(8.85 \times 10^{-14} F / cm\right)}{9 \times 10^{-7} cm} = 220\mu A / V^2 \cdot$$

The device transconductance parameters are equal:

$$K_P = k'_P \left(\frac{W_P}{L_P}\right) = 88\mu A / V^2 \left(\frac{3.0\mu m}{0.6\mu m}\right) = 440\mu A / V^2$$

and

$$K_N = k'_N \left(\frac{W_N}{L_N}\right) = 220\mu A / V^2 \left(\frac{1.2\mu m}{0.6\mu m}\right) = 440\mu A / V^2 .$$

Using the detailed expression the propagation delay is

$$\begin{aligned} t_P &= \frac{C_L}{K\left(V_{DD} - V_T\right)}\left[\frac{2V_T}{\left(V_{DD} - V_T\right)} + \ln\left(\frac{3V_{DD} - 4V_T}{V_{DD}}\right)\right] \\ &= \frac{1 \times 10^{-12} F}{\left(440 \times 10^{-6} A / V^2\right)(2.5V - 0.5V)}\left[\frac{2(0.5V)}{(2.5V - 0.5V)} + \ln\left(\frac{3(2.5V) - 4(0.5V)}{2.5V}\right)\right] \cdot \\ &= 1.45ns \end{aligned}$$

Using the approximate expression,

$$t_P \approx \frac{1.6C_L}{K\left(V_{DD} - V_T\right)} = \frac{1.6\left(10^{-12} F\right)}{\left(440 \times 10^{-6} A / V^2\right)(2.5V - 0.5V)} = 1.8ns ,$$

which is 24% higher than the value obtained using the detailed equation.

Example 6.6 Propagation Delays for the Minimum-Size CMOS Inverter

Estimate t_{PLH} and t_{PHL} for the minimum-size CMOS inverter of Figure 6.24 with V_{DD} = 2.5 V, V_T = 0.5 V, and t_{OX} = 9 nm using the detailed and approximate equations, assuming a lumped 1 pF load.

Solution: The process transconductance values are $k'_P = 88\mu A / V^2$ and $k'_N = 220\mu A / V^2$. In the minimum-size inverter, all MOSFET gate dimensions are set to the minimum value, $W_N = 2L_N = W_P = 2L_P$, so that the device transconductance parameters are twice the process transconductance parameters: $K_P = 2k'_P$ and $K_N = 2k'_N$.

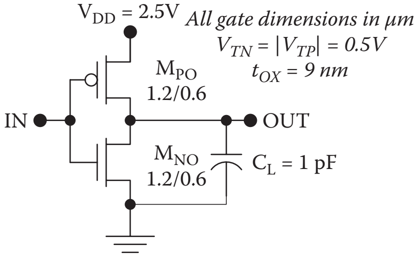


FIGURE 6.24
Example minimum-size inverter for the calculation of the propagation delays.

The low-to-high propagation delay is

$$t_{PLH} = \frac{C_L}{K_P\left(V_{DD}+V_{TP}\right)}\left[\frac{-2V_{TP}}{\left(V_{DD}+V_{TP}\right)}+\ln\left(\frac{3V_{DD}+4V_{TP}}{V_{DD}}\right)\right]$$

$$= \frac{1\times 10^{-12}F}{\left(176\times 10^{-6}A/V^2\right)(2.5V-0.5V)}\left[\frac{2(0.5V)}{(2.5V-0.5V)}+\ln\left(\frac{3(2.5V)-4(0.5V)}{2.5V}\right)\right]$$

$$= 3.6ns,$$

but the high-to-low propagation delay is

$$t_{PHL} = \frac{C_L}{K_N\left(V_{DD}-V_{TN}\right)}\left[\frac{2V_{TN}}{\left(V_{DD}-V_{TN}\right)}+\ln\left(\frac{3V_{DD}-4V_{TN}}{V_{DD}}\right)\right]w$$

$$= \frac{1\times 10^{-12}F}{\left(440\times 10^{-6}A/V^2\right)(2.5V-0.5V)}\left[\frac{2(0.5V)}{(2.5V-0.5V)}+\ln\left(\frac{3(2.5V)-4(0.5V)}{2.5V}\right)\right]$$

$$= 1.45ns.$$

Because the p-MOS transistor is weaker, t_{PHL} is about 2.5 times longer than t_{PLH}.

6.6.6 Effect of the Input Rise and Fall Time

In the previous sections, we derived expressions for the propagation delay times with the assumption of abrupt voltage transitions at the input to the circuit. In a real digital system, the input signals will have finite rise and fall times that will tend to increase the propagation delays. Consider first the high-to-low propagation delay with a finite rise time* at the input as shown in Figure 6.25. There are two consequences of the finite input rise time: (1) the

* The rise time is defined as the time required for the input signal to increase from the 10% point to the 90% point.

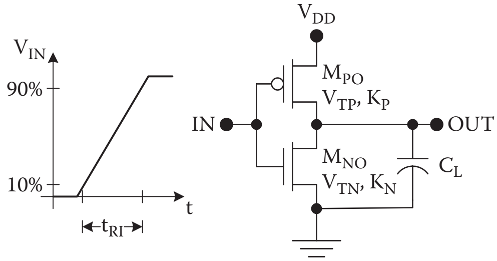


FIGURE 6.25
CMOS inverter with an input signal having a finite rise time t_R.

n-MOS transistor does not turn on abruptly but instead exhibits a gradual rise in its drain current; and (2) the p-MOS device does not turn off abruptly, resulting in some crossover current that loads the n-MOS pull-down device. Both of these effects tend to increase the propagation delay for the inverter circuit, although it is measured from the (delayed) 50% point on the input waveform.

Because of the gradual turn-on of the pull-down device and the gradual turn-off of the pull-up device, the delay time analysis is more complicated than the derivations given in the previous sections. However, the high-to-low propagation delay may be estimated using the empirical relationship

$$t'_{PHL} \approx \sqrt{t_{PHL}^2 + t_{RI}^2 / 2}\,, \tag{6.57}$$

where t_{PHL} is the intrinsic delay for the gate with abrupt input transitions, t_{RI} is the rise time for the input signal, and t'_{PHL} is the extrinsic delay for the gate with a non-abrupt input transition. Figure 6.26 shows the extrinsic delay calculated using the empirical relationship of Equation 6.57 as well as the extrinsic delay determined using SPICE with the assumption of a trapezoidal input waveform. Real input waveforms are usually nontrapezoidal, and this accentuates the dependence of the extrinsic delay on the rise time.

Following the same line of reasoning, a finite fall time at the input as shown in Figure 6.27 will increase t_{PLH}.

The approximate value of the extrinsic propagation delay in this case is

$$t'_{PLH} \approx \sqrt{t_{PLH}^2 + t_{FI}^2 / 2}\,, \tag{6.58}$$

where t_{PLH} is the intrinsic delay for the gate with an abrupt input transition, and t_{FI} is the input fall time.

It is important to note that the preceding relationships are very approximate, and as discussed above, the extrinsic propagation delays depend on the *shape* of the input waveform as well as its rise and fall times.

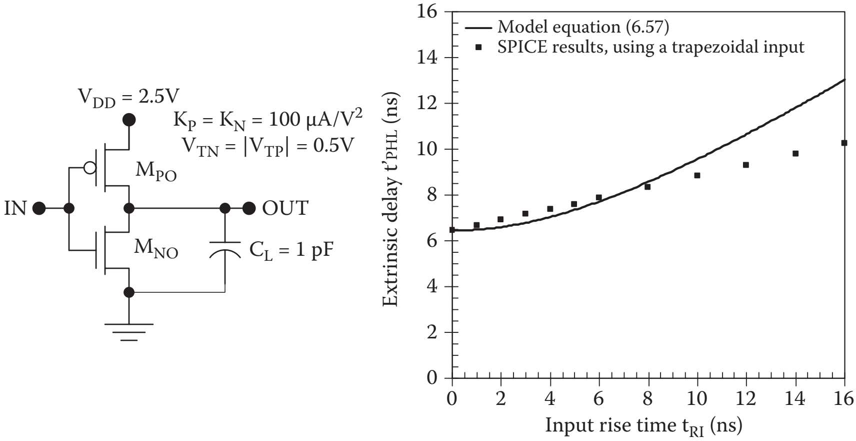


FIGURE 6.26
Extrinsic high-to-low propagation delay for a symmetric CMOS gate as a function of the input rise time.

6.7 Inverter Rise and Fall Times

The output rise time for a gate circuit is of importance because, as is clear from the previous section, it will affect the propagation delay times for the fan-out gates.

6.7.1 Fall Time

If we return to our assumption of an abrupt input transition, the calculation of the fall time for a CMOS inverter follows along the same lines as the t_{PHL} calculation presented previously. Thus, if the input waveform makes

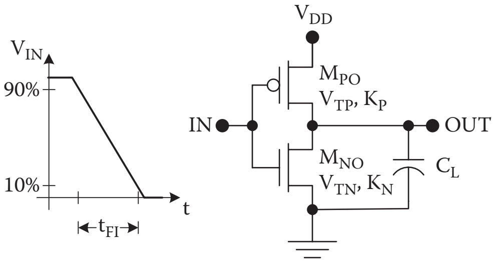


FIGURE 6.27
CMOS inverter with an input signal having a finite fall time t_F.

an abrupt transition from 0 to V_{DD} at t = 0 and the load is a lumped capacitance as shown in Figure 6.28, the fall time is the sum of two components $t_F = t_{F1} + t_{F2}$, where t_{F1} and t_{F2} represent the time intervals for saturated and linear operation of the n-MOS transistor, respectively.

It has already been shown that the n-MOS transistor operates in the saturation region until V_{OUT} drops to $V_{DD} - V_{TN}$, and the length of this time interval is t_{F1}:

$$t_{F1} = \frac{2V_{TN}C_L}{K_N\left(V_{DD} - V_{TN}\right)^2} \ . \tag{6.59}$$

After V_{OUT} drops below $V_{DD} - V_T$, the n-MOS operates in the linear mode until the end of the fall time, which corresponds to $V_{OUT} = V_{DD}/10$ (the 10% point). The length of this time interval is t_{F2}, given by

$$\begin{aligned} t_{F2} &= -C_L \int_{V_{DD}-V_{TN}}^{V_{DD}/10} \frac{dV_{OUT}}{K_N\left[\left(V_{DD} - V_{TN}\right)V_{OUT} - V_{OUT}^2 / 2\right]} \\ &= -\frac{C_L}{K_N\left(V_{DD} - V_{TN}\right)}\left[ln\left(\frac{V_{OUT}}{V_{OUT} - 2\left(V_{DD} - V_{TN}\right)}\right)\right]\Bigg|_{V_{DD}-V_{TN}}^{V_{DD}/10} \\ &= \frac{C_L}{K_N\left(V_{DD} - V_{TN}\right)}\ln\left(\frac{19V_{DD} - 20V_{TN}}{V_{DD}}\right). \end{aligned} \tag{6.60}$$

The fall time is given by the sum $t_{F1} + t_{F2}$, so that

$$t_F = \frac{C_L}{K_N\left(V_{DD} - V_{TN}\right)}\left[\frac{2V_{TN}}{\left(V_{DD} - V_{TN}\right)} + \ln\left(\frac{19V_{DD} - 20V_{TN}}{V_{DD}}\right)\right]. \tag{6.61}$$

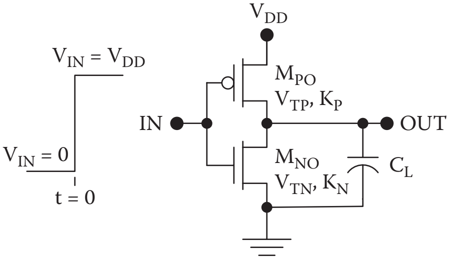


FIGURE 6.28
CMOS inverter with a lumped capacitive load and an abrupt low-to-high voltage transition at the input for the estimation of t_F.

This expression is similar to that for the high-to-low propagation delay time, and the fall time is proportional to C_L but inversely proportional to K_N.

6.7.2 Rise Time

The rise time may be determined in similar manner, with the assumption of an abrupt transition at the input, yielding

$$t_R = \frac{C_L}{K_P\left(V_{DD}+V_{TP}\right)}\left[\frac{-2V_{TP}}{\left(V_{DD}+V_{TP}\right)}+\ln\left(\frac{19V_{DD}+20V_{TP}}{V_{DD}}\right)\right]. \tag{6.62}$$

Example 6.7 Rise and Fall Times for the Symmetric CMOS Inverter

Estimate t_R and t_F for the symmetric CMOS inverter depicted in Figure 6.29 with $V_{DD} = 2.5$ V, $V_{TN} = |V_{TP}| = 0.5$ V, and $t_{OX} = 9$ nm, assuming that the input voltage makes abrupt transitions.

Solution: In this symmetric inverter, the device transconductance parameters are equal:

$$\begin{aligned} K_P = K_N &= \frac{W_N}{L_N}\frac{\mu_n\varepsilon_{OX}}{t_{OX}} = \left(\frac{1.2\mu m}{0.6\mu m}\right)\frac{\left(580cm^2/Vs\right)(3.9)\left(8.85\times10^{-14}F/cm\right)}{9\times10^{-7}cm} \\ &= 440\mu A/V^2. \end{aligned}$$

The rise time is

$$\begin{aligned} t_R &= \frac{C_L}{K\left(V_{DD}+V_{TP}\right)}\left[\frac{-2V_{TP}}{\left(V_{DD}+V_{TP}\right)}+\ln\left(\frac{19V_{DD}+20V_{TP}}{V_{DD}}\right)\right] \\ &= \frac{1\times10^{-12}F}{\left(440\times10^{-6}A/V^2\right)\left(2.5V-0.5V\right)}\left[\frac{2(0.5V)}{(2.5V-0.5V)}+\ln\left(\frac{19(2.5V)-20(0.5V)}{2.5V}\right)\right] \\ &= 3.6ns \end{aligned}$$

and the fall time is

$$\begin{aligned} t_F &= \frac{C_L}{K\left(V_{DD}-V_{TN}\right)}\left[\frac{2V_{TN}}{\left(V_{DD}-V_{TN}\right)}+\ln\left(\frac{19V_{DD}-20V_{TN}}{V_{DD}}\right)\right] \\ &= \frac{1\times10^{-12}F}{\left(440\times10^{-6}A/V^2\right)\left(2.5V-0.5V\right)}\left[\frac{2(0.5V)}{(2.5V-0.5V)}+\ln\left(\frac{19(2.5V)-20(0.5V)}{2.5V}\right)\right] \\ &= 3.6ns. \end{aligned}$$

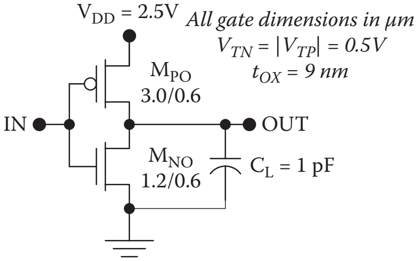


FIGURE 6.29
Example symmetric inverter for the calculation of the rise and fall times.

As expected for the symmetric circuit, the rise and fall times are equal, and they are 2.5 times the propagation delays.

Example 6.8 Rise and Fall Time for the Minimum-Size CMOS Inverter

Estimate t_R and t_F for the minimum-size CMOS inverter of Figure 6.30 with $V_{DD} = 2.5$ V, $V_{TN} = |V_{TP}| = 0.5$ V, and $t_{OX} = 9$ nm.

Solution: Here,

$$K_N = \frac{W_N}{L_N}\frac{\mu_n \varepsilon_{OX}}{t_{OX}} = \left(\frac{1.2\mu m}{0.6\mu m}\right)\frac{\left(580cm^2/Vs\right)(3.9)\left(8.85\times10^{-14}F/cm\right)}{9\times10^{-7}cm} = 440\mu A/V^2$$

but

$$K_P = K_N\left(\frac{\mu_p}{\mu_n}\right) \approx K_N/2.5\ .$$

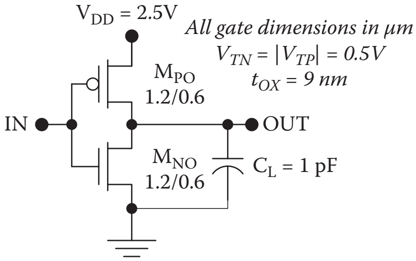


FIGURE 6.30
Example minimum-size inverter for the calculation of the rise and fall times.

The fall time is the same as before:

$$t_F = \frac{C_L}{K\left(V_{DD} - V_{TN}\right)}\left[\frac{2V_{TN}}{\left(V_{DD} - V_{TN}\right)} + \ln\left(\frac{19V_{DD} - 20V_{TN}}{V_{DD}}\right)\right]$$

$$= \frac{1\times 10^{-12}F}{\left(440\times 10^{-6}A/V^2\right)\left(2.5V - 0.5V\right)}\left[\frac{2\left(0.5V\right)}{\left(2.5V - 0.5V\right)} + \ln\left(\frac{19\left(2.5V\right) - 20\left(0.5V\right)}{2.5V}\right)\right]$$

$$= 3.6ns$$

but because of the weaker p-MOS transistor, the rise time is

$$t_R \approx 2.5t_F = 9.1ns\ .$$

6.7.3 Effect of the Input Rise and Fall Time on Output Rise and Fall Time

In the previous sections, the rise and fall time were derived with the simplifying assumption of an abrupt input voltage transition. However, a finite transition time at the input of a CMOS inverter will increase the rise and fall time at the output, for the same reasons that the propagation delays are increased. For the case of a trapezoidal input waveform (meaning both the low-to-high and high-to-low transitions are linear), it can be shown that the extrinsic rise and fall times t'_R and t'_F are given by

$$t'_R = \sqrt{t_R^2 + t_{FI}^2/7} \tag{6.63}$$

and

$$t'_F = \sqrt{t_F^2 + t_{RI}^2/7}\ , \tag{6.64}$$

where t_R and t_F are the intrinsic rise and fall times, calculated for the case of abrupt input transitions, and t_{RI} and t_{FI} are the rise time and fall time, respectively, for the input signal.

Example 6.9 Inverter Design with Delay Constraints

Design a CMOS inverter such that the switching threshold is $V_{DD}/2$, $t_{PLH} \le 250ps$, and $t_{PHL} \le 200ps$ with, $C_L = 500fF$. $V_{DD} = 2.5V$, $V_{TN} = \left|V_{TP}\right| = 0.5V$; the minimum gate dimension (L or W) is $0.6\mu m$, and the oxide thickness is 9 nm.

Solution: The switching threshold requirement sets the transconductance ratio $K_R = K_N / K_P$:

$$K_R = \left(\frac{V_{DD} + V_{TP} - V_M}{V_M - V_{TN}}\right)^2 = \left(\frac{2.5V - 0.5V - 1.25V}{1.25V - 0.5V}\right)^2 = 1.0\ .$$

The delay factors are equal because of the symmetry in the threshold voltages:

$$\Gamma_P = \Gamma_N = \frac{1}{(V_{DD} - V_{TN})}\left[\frac{2V_{TN}}{(V_{DD} - V_{TN})} + \ln\left(\frac{3V_{DD} - 4V_{TN}}{V_{DD}}\right)\right]$$

$$= \frac{1}{(2.5V - 0.5V)}\left[\frac{2(0.5V)}{(2.5V - 0.5V)} + \ln\left(\frac{3(2.5V) - 4(0.5V)}{2.5V}\right)\right] = 0.644V^{-1}.$$

The requirements on the individual device transconductance parameters are

$$K_P \geq \frac{C_L \Gamma_P}{t_{PLH,\max}} = \frac{(500\times 10^{-15}F)(0.644V^{-1})}{(250\times 10^{-12}s)} = 1.29mA/V^2$$

and

$$K_N \geq \frac{C_L \Gamma_N}{t_{PHL,\max}} = \frac{(500\times 10^{-15}F)(0.644V^{-1})}{(200\times 10^{-12}s)} = 1.61mA/V^2 \cdot$$

The process transconductance values for the p-MOS and n-MOS transistors are

$$k'_P = \frac{\mu_p \varepsilon_{OX}}{t_{OX}} = \frac{(230cm^2/Vs)(3.9)(8.85\times 10^{-14}F/cm)}{9\times 10^{-7}cm} = 88\mu A/V^2$$

and

$$k'_N = \frac{\mu_n \varepsilon_{OX}}{t_{OX}} = \frac{(580cm^2/Vs)(3.9)(8.85\times 10^{-14}F/cm)}{9\times 10^{-7}cm} = 220\mu A/V^2 \cdot$$

Therefore, the required aspect ratios are

$$\frac{W_P}{L_P} \geq \frac{K_{P,\min}}{k'_P} = \frac{1290\mu A/V^2}{88\mu A/V^2} = 14.7$$

and

$$\frac{W_N}{L_N} \geq \frac{K_{N,\min}}{k'_N} = \frac{1610\mu A/V^2}{220\mu A/V^2} = 7.3 \cdot$$

From the requirement on the switching threshold, we also have

$$\frac{W_N/L_N}{W_P/L_P} = K_R(\mu_p/\mu_n) \approx 1/2.5.$$

If we fix the gate lengths at the minimum, $L_P = L_N = 0.6\mu m$, then using a 25% safety margin the gate lengths may be chosen as

$$W_N = 7.3(1+25\%)0.6\mu m = 5.5\mu m$$

and

$$W_P = 2.5W_N = 13.8\mu m \cdot$$

Note that W_P meets the t_{PLH} requirement with better than a 25% safety margin, because $W_P / L_P = 13.8\mu m / 0.6\mu m = 23$.

6.8 Propagation Delays in Short-Channel CMOS

In the previous sections, gate delays and rise/fall times were derived for CMOS circuits using the long-channel MOSFET equations. As shown in Chapter 4, it is necessary to use a different set of equations for short-channel MOS transistors because of the field-dependent mobilities and velocity saturation. Here we present approximate propagations delay and design equations based these short-channel MOS equations.

6.8.1 High-to-Low Propagation Delay t_{PHL} in Short-Channel CMOS

Consider a short-channel CMOS inverter with a lumped capacitive load and an abrupt low-to-high input transition as shown in Figure 6.31.

At $t = 0^+$, the n-MOS transistor becomes saturated and the p-MOS transistor becomes cutoff. As shown in Chapter 4, a short-channel MOS transistor saturates at a lower drain-to-source voltage than a long-channel device with the same threshold voltage and gate-to-source bias. For this approximate analysis,

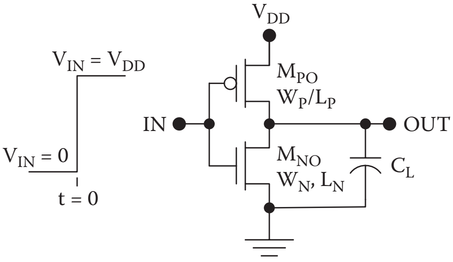


FIGURE 6.31
CMOS inverter with a lumped capacitive load and an abrupt low-to-high voltage transition at the input for the estimation of t_{PHL}.

we will make the assumption that the n-MOS transistor remains saturated throughout the propagation delay, with a saturated current given by

$$I_{DSATn} = C_{ox} W_N v_{satn} \left(V_{DD} - V_{TN}\right) \frac{\sqrt{1 + 2\mu_n \left(V_{DD} - V_{TN}\right) / \left(v_{satn} L_N\right)} - 1}{\sqrt{1 + 2\mu_n \left(V_{DD} - V_{TN}\right) / \left(v_{satn} L_N\right)} + 1}, \quad (6.65)$$

where v_{satn} is the saturation velocity for electrons (approximately $9 \times 10^6 cm/s$ in Si). The high-to-low propagation delay is then approximately

$$t_{PHL} \approx \frac{C_L V_{DD}}{2 I_{DSATn}} = \frac{C_L V_{DD}}{2 C_{ox} W_N v_{satn} \left(V_{DD} - V_{TN}\right)} \frac{\sqrt{1 + 2\mu_n \left(V_{DD} - V_{TN}\right) / \left(v_{satn} L_N\right)} + 1}{\sqrt{1 + 2\mu_n \left(V_{DD} - V_{TN}\right) / \left(v_{satn} L_N\right)} - 1}. \quad (6.66)$$

6.8.2 Low-to-High Propagation Delay t_{PLH} in Short-Channel CMOS

Consider a short-channel CMOS inverter with a lumped capacitive load and an abrupt high-to-low input transition as shown in Figure 6.32.

At $t = 0^+$, the p-MOS transistor becomes saturated and the n-MOS transistor becomes cutoff. We will make the assumption that the p-MOS transistor remains saturated throughout the propagation delay with a saturated current:

$$I_{DSATp} = C_{ox} W_P v_{satp} \left(V_{DD} + V_{TP}\right) \frac{\sqrt{1 + 2\mu_p \left(V_{DD} + V_{TP}\right) / \left(v_{satp} L_P\right)} - 1}{\sqrt{1 + 2\mu_p \left(V_{DD} + V_{TP}\right) / \left(v_{satp} L_P\right)} + 1}, \quad (6.67)$$

where v_{satp} is the sa)turation velocity for holes (approximately $8 \times 10^6 cm/s$ in Si). The high-to-low propagation delay is then approximately

$$t_{PLH} \approx \frac{C_L V_{DD}}{2 I_{DSATp}} = \frac{C_L V_{DD}}{2 C_{ox} W_P v_{satp} \left(V_{DD} + V_{TP}\right)} \frac{\sqrt{1 + 2\mu_p \left(V_{DD} - V_{TP}\right) / \left(v_{satp} L_P\right)} + 1}{\sqrt{1 + 2\mu_p \left(V_{DD} - V_{TP}\right) / \left(v_{satp} L_P\right)} - 1}. \quad (6.68)$$

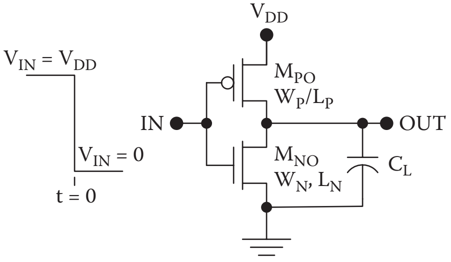


FIGURE 6.32
CMOS inverter with a lumped capacitive load and an abrupt low-to-high voltage transition at the input for the estimation of t_{PHL}.

6.8.3 Comparison of the Short-Channel and Long-Channel Delay Equations

Based on long-channel MOSFET equations, we derived the inverter high-to-low propagation delay to be

$$t_{PHL} = \frac{C_L}{\mu_n C_{ox}\left(W_N / L_N\right)\left(V_{DD} - V_{TN}\right)}\left[\frac{2V_{TN}}{\left(V_{DD} - V_{TN}\right)} + \ln\left(\frac{3V_{DD} - 4V_{TN}}{V_{DD}}\right)\right]$$

$$= \frac{C_L}{\mu_n C_{ox}\left(W_N / L_N\right)}\Gamma_n. \tag{6.69}$$

The high-to-low delay based on short-channel behavior may be written as

$$t_{PHL} \approx \frac{C_L V_{DD}}{2C_{ox} W_N v_{satn}\left(V_{DD} - V_{TN}\right)} \frac{\sqrt{1 + 2\mu_n\left(V_{DD} - V_{TN}\right)/\left(v_{satn} L_N\right)} + 1}{\sqrt{1 + 2\mu_n\left(V_{DD} - V_{TN}\right)/\left(v_{satn} L_N\right)} - 1}$$

$$= \frac{C_L}{\mu_n C_{ox}\left(W_N / L_N\right)}\Gamma_{neff}. \tag{6.70}$$

where

$$\Gamma_{neff} \approx \frac{\mu_n V_{DD}}{2v_{satn} L_N\left(V_{DD} - V_{TN}\right)} \frac{\sqrt{1 + 2\mu_n\left(V_{DD} - V_{TN}\right)/\left(v_{satn} L_N\right)} + 1}{\sqrt{1 + 2\mu_n\left(V_{DD} - V_{TN}\right)/\left(v_{satn} L_N\right)} - 1}. \tag{6.71}$$

Thus, the ratio Γ_{neff} / Γ_n gives the approximate ratio of the actual delay divided by the long-channel value.

Similarly for the low-to-high propagation delay, from the long-channel MOSFET equations, we obtain

$$t_{PLH} = \frac{C_L}{\mu_p C_{ox}\left(W_P / L_P\right)\left(V_{DD} + V_{TP}\right)}\left[\frac{2V_{TP}}{\left(V_{DD} + V_{TP}\right)} + \ln\left(\frac{3V_{DD} + 4V_{TP}}{V_{DD}}\right)\right]$$

$$= \frac{C_L}{\mu_p C_{ox}\left(W_P / L_P\right)}\Gamma_p, \tag{6.72}$$

and from the short-channel equations, we have

$$t_{PLH} \approx \frac{C_L V_{DD}}{2C_{ox} W_P v_{satp}\left(V_{DD} + V_{TP}\right)} \frac{\sqrt{1 + 2\mu_p\left(V_{DD} + V_{TP}\right)/\left(v_{satp} L_P\right)} + 1}{\sqrt{1 + 2\mu_p\left(V_{DD} - V_{TP}\right)/\left(v_{satp} L_P\right)} - 1}$$

$$= \frac{C_L}{\mu_p C_{ox}\left(W_P / L_P\right)}\Gamma_{peff}. \tag{6.73}$$

where

$$\Gamma_{peff} \approx \frac{\mu_p V_{DD}}{2v_{satp}L_P\left(V_{DD}+V_{TP}\right)}\frac{\sqrt{1+2\mu_p\left(V_{DD}+V_{TP}\right)/\left(v_{satp}L_P\right)}+1}{\sqrt{1+2\mu_p\left(V_{DD}+V_{TP}\right)/\left(v_{satp}L_P\right)}-1}. \tag{6.74}$$

Here the ratio Γ_{peff}/Γ_p gives the approximate ratio of the actual delay divided by the long-channel value.

6.8.4 Propagation Delay Design Equations for Short-Channel CMOS

By rearranging Equations 6.70 and 6.73, we can develop equations that allow us to choose the transistors' widths to meet a delay constraint. If the propagation delays should be less than some value $t_{P,\max}$ with an external load capacitance C_L, then for an inverter the transistor widths should be chosen according to

$$W_N \geq \frac{C_L V_{DD}}{2C_{ox}t_{P,\max}v_{satn}\left(V_{DD}-V_{TN}\right)}\frac{\sqrt{1+2\mu_n\left(V_{DD}-V_{TN}\right)/\left(v_{satn}L_N\right)}+1}{\sqrt{1+2\mu_n\left(V_{DD}-V_{TN}\right)/\left(v_{satn}L_N\right)}-1} \tag{6.75}$$

and

$$W_P \geq \frac{C_L V_{DD}}{2C_{ox}t_{P,\max}v_{satp}\left(V_{DD}+V_{TP}\right)}\frac{\sqrt{1+2\mu_p\left(V_{DD}+V_{TP}\right)/\left(v_{satp}L_P\right)}+1}{\sqrt{1+2\mu_p\left(V_{DD}+V_{TP}\right)/\left(v_{satp}L_P\right)}-1}. \tag{6.76}$$

Example 6.10 Propagation Delays in Short-Channel CMOS

Estimate the propagation delays for the CMOS circuit of Figure 6.33, assuming abrupt input transitions with rail-to-rail swing.

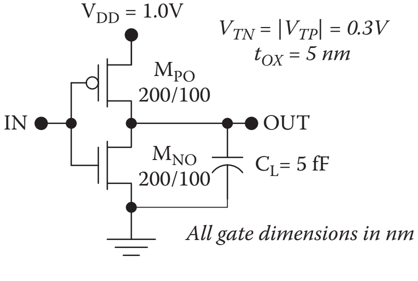


FIGURE 6.33
Example short-channel CMOS inverter for the calculation of the propagation delays.

Solution: The oxide capacitance per unit area is

$$C_{ox}=\frac{\varepsilon_{ox}}{t_{ox}}=\frac{3.9\left(8.85\times10^{-14}F/cm\right)}{5\times10^{-7}cm}=6.9\times10^{-7}F/cm^2\cdot$$

Based on the short-channel MOSFET equations,

$$I_{DSATn}=C_{ox}W_Nv_{satn}\left(V_{DD}-V_{TN}\right)\frac{\sqrt{1+2\mu_n\left(V_{DD}-V_{TN}\right)/\left(v_{satn}L_N\right)}-1}{\sqrt{1+2\mu_n\left(V_{DD}-V_{TN}\right)/\left(v_{satn}L_N\right)}+1}=87\mu A$$

and

$$t_{PHL}\approx\frac{C_LV_{DD}}{2I_{DSATn}}=\frac{\left(5\times10^{-15}F\right)\left(1.0V\right)}{2\left(87\times10^{-6}A\right)}=29ps\,;$$

$$I_{DSATp}=C_{ox}W_Pv_{satp}\left(V_{DD}+V_{TP}\right)\frac{\sqrt{1+2\mu_p\left(V_{DD}+V_{TP}\right)/\left(v_{satp}L_P\right)}-1}{\sqrt{1+2\mu_p\left(V_{DD}+V_{TP}\right)/\left(v_{satp}L_P\right)}+1}=77\mu A\,,$$

and

$$t_{PLH}\approx\frac{C_LV_{DD}}{2I_{DSATp}}=\frac{\left(5\times10^{-15}F\right)\left(1.0V\right)}{2\left(77\times10^{-6}A\right)}=32ps\ .$$

The propagation delays are similar in value because the carrier saturation velocities are close in value.

Example 6.11 Comparison of Short-Channel and Long-Channel CMOS Delays

For CMOS inverters with $V_{DD}=1.0V$, $V_{TN}=\left|V_{TP}\right|=0.3V$, $W_N=W_P=10\mu m$, and $C_L=1pF$, calculate Γ_{neff}/Γ_n as a function of the gate length and hence determine the ratio by which the high-to-low propagation delay is underestimated based on the long-channel MOSFET equations. Repeat for Γ_{peff}/Γ_p and the low-to-high propagation delay.

Solution: The ratio Γ_{neff}/Γ_n is

$$\frac{t_{PHL}}{t_{PHL}\left(long\,channel\right)}=\frac{\Gamma_{neff}}{\Gamma_n}\approx\frac{\left[\dfrac{\mu_nV_{DD}}{2v_{satn}L_N}\,\dfrac{\sqrt{1+2\mu_n\left(V_{DD}-V_{TN}\right)/\left(v_{satn}L_N\right)}+1}{\sqrt{1+2\mu_n\left(V_{DD}-V_{TN}\right)/\left(v_{satn}L_N\right)}-1}\right]}{\left[\dfrac{2V_{TN}}{\left(V_{DD}-V_{TN}\right)}+\ln\left(\dfrac{3V_{DD}-4V_{TN}}{V_{DD}}\right)\right]},$$

and the ratio Γ_{peff} / Γ_p is

$$\frac{t_{PLH}}{t_{PLH}\left(long\,channel\right)} = \frac{\Gamma_{peff}}{\Gamma_p} \approx \frac{\left[\dfrac{\mu_p V_{DD}}{2v_{satp}L_P}\dfrac{\sqrt{1+2\mu_p\left(V_{DD}+V_{TP}\right)/\left(v_{satp}L_P\right)}+1}{\sqrt{1+2\mu_p\left(V_{DD}+V_{TP}\right)/\left(v_{satp}L_P\right)}-1}\right]}{\left[\dfrac{2V_{TP}}{\left(V_{DD}+V_{TP}\right)}+\ln\left(\dfrac{3V_{DD}+4V_{TP}}{V_{DD}}\right)\right]}.$$

These ratios are plotted as functions of the gate lengths L_N and L_P in Figures 6.34 and 6.35. The equations based on long-channel transistors underestimate t_{PHL} (t_{PLH}) by a factor of ¼ (1/2.5) with 100 nm gate lengths.

Example 6.12 Design of Short-Channel CMOS with Delay Constraints

Choose the widths of the transistors in a CMOS inverter such that $t_P \le 50ps$ with $C_L = 25fF$. $V_{DD} = 1.0V$ and $V_{TN} = |V_{TP}| = 0.3V$. $L_P = L_N = 100nm$ and $t_{ox} = 5nm$.

Solution: The requirements on the device widths are

$$W_P \ge \frac{C_L V_{DD}}{2C_{ox}t_{P,\max}v_{satp}\left(V_{DD}+V_{TP}\right)}\frac{\sqrt{1+2\mu_p\left(V_{DD}+V_{TP}\right)/\left(v_{satp}L_P\right)}+1}{\sqrt{1+2\mu_p\left(V_{DD}+V_{TP}\right)/\left(v_{satp}L_P\right)}-1} = 1.7\mu m$$

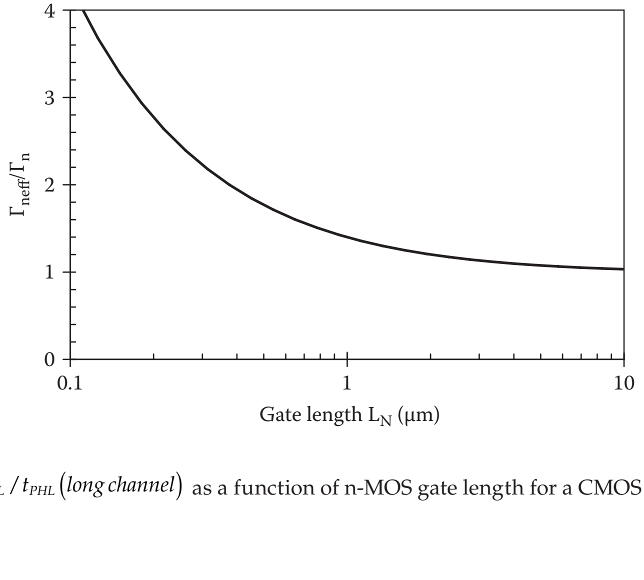


FIGURE 6.34
$\Gamma_{neff} / \Gamma_n = t_{PHL} / t_{PHL}\left(long\,channel\right)$ as a function of n-MOS gate length for a CMOS inverter.

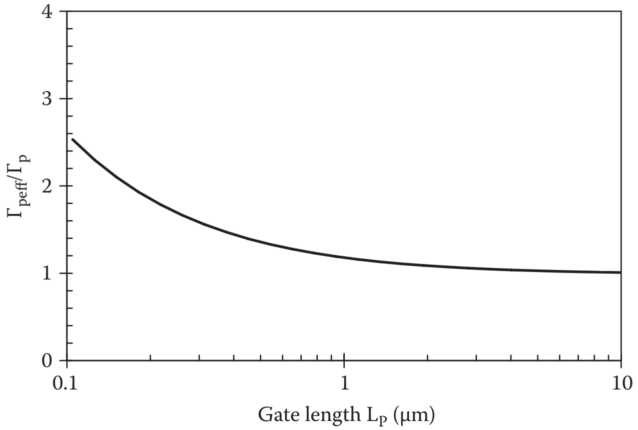


FIGURE 6.35
$\Gamma_{peff}/\Gamma_p = t_{PLH}/t_{PLH}(long\,channel)$ as a function of p-MOS gate length for a CMOS inverter.

and

$$W_N \geq \frac{C_L V_{DD}}{2C_{ox} t_{n,\max} v_{satn}(V_{DD}-V_{TN})} \frac{\sqrt{1+2\mu_n(V_{DD}-V_{TN})/(v_{satn}L_N)}+1}{\sqrt{1+2\mu_n(V_{DD}-V_{TN})/(v_{satn}L_N)}-1} = 1.1\mu m\,.$$

In a short-channel CMOS inverter, a transistor size ratio of $W_P/W_N = 2.5$ is not necessary for symmetric delay performance.

6.9 Power Dissipation

Broadly speaking, the dissipation in CMOS can be classified as static (DC) dissipation and dynamic (AC) dissipation. The total dissipation is the sum of these two components:

$$P = P_{DC} + P_{AC}\,. \tag{6.77}$$

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. Thus,

$$P = \underbrace{P_{subthreshold} + P_{pn}}_{P_{DC}} + \underbrace{P_{sc} + P_{switch}}_{P_{AC}}, \tag{6.78}$$

where $P_{subthreshold}$ is power associated with MOSFET subthreshold conduction, P_{pn} is power associated with p-n junction leakage in the MOSFETs, P_{sc} is short-circuit dissipation, and P_{switch} is capacitance switching dissipation.

Usually, the capacitance switching power is dominant in CMOS circuitry. However, all four components should be considered in low-power VLSI design.

6.9.1 Capacitance Switching Dissipation

Suppose a CMOS gate is loaded by a lumped capacitance as shown in Figure 6.36, 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

$$J_{switch} = V_{DD}C_L \int_0^{V_{DD}} dV = C_L V_{DD}^2 . \tag{6.79}$$

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

$$P_{switch} = fC_L V_{DD}^2 . \tag{6.80}$$

Normally, the output node of a CMOS gate switches at a frequency that is lower than the system clock frequency. This is accounted for by introducing an activity factor α :

$$P_{switch} = \alpha f_{CLK} C_L V_{DD}^2 . \tag{6.81}$$

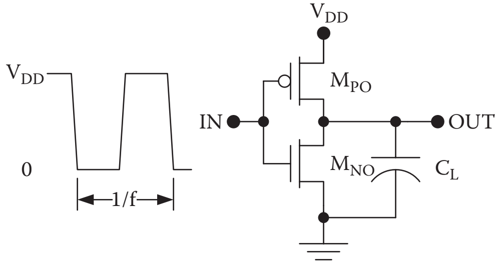


FIGURE 6.36
CMOS inverter for the consideration of the capacitance switching power.

The capacitance that loads the output node comprises three important components:

$$C_L = C_{out} + C_{\text{interconnect}} + NC_{in}\,, \tag{6.82}$$

where C_{out} is the output capacitance of the gate circuit, $C_{\text{interconnect}}$ is the capacitance of the interconnect wires, and NC_{in} is the load capacitance associated with N fan-out gates.

In complex CMOS logic gates, there are internal nodes that may switch even when the output node does not. To accurately account for all of the capacitance switching power, it is necessary to sum up the contributions from all of the individual nodes:

$$P_{switch} = f_{CLK} V_{DD} \sum_{i=1}^{N} \alpha_{\text{i}} C_i V_i\,, \tag{6.83}$$

where α_i is the activity for the ith node, C_i is the capacitance loading the ith node, and V_i is the average voltage swing on the ith node.

6.9.2 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, whereas 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.

Consider a symmetric CMOS inverter with a negligible capacitive load as shown in Figure 6.37.

For $V_{IN} \le V_{DD}/2$, the n-MOSFET is saturated, and for $V_{IN} \ge V_{DD}/2$, the p-MOSFET is saturated. The supply current as a function of the input voltage is

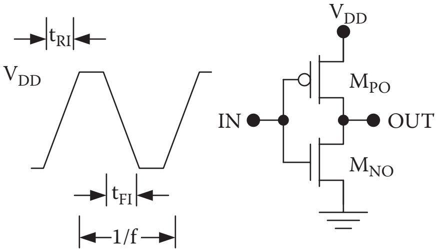


FIGURE 6.37
CMOS inverter for the consideration of the short-circuit power.

$$i_{DD} = \begin{cases} 0; & V_{IN} \le V_T \\ \dfrac{K}{2}\left(V_{IN} - V_T\right)^2; & V_{IN} \le V_{DD}/2 \\ \dfrac{K}{2}\left(V_{DD} - V_{IN} - V_T\right)^2; & V_{IN} \ge V_{DD}/2 \\ 0; & V_{IN} \ge V_{DD} - V_T \end{cases} . \tag{6.84}$$

Suppose the rise and fall times for the input waveform are equal:

$$t_{RI} = t_{FI} = \tau . \tag{6.85}$$

Then the time-averaged short-circuit power is

$$\begin{aligned} P_{sc} &= \frac{V_{DD}}{T} \int_0^T i_{DD} dt \\ &= \frac{2V_{DD}}{T} \left\{ \int_{\tau V_T / V_{DD}}^{\tau/2} \frac{K}{2} \left(\frac{V_{DD} t}{\tau} - V_T \right)^2 dt + \int_{\tau/2}^{\tau - \tau V_T / V_{DD}} \frac{K}{2} \left(V_{DD} - \frac{V_{DD} t}{\tau} - V_T \right)^2 dt \right\} \\ &= \frac{K \tau f \left(V_{DD} - 2V_T\right)^3}{12} . \end{aligned} \tag{6.86}$$

If the switching frequency is less than the clock frequency, we can account for this by invoking the switching activity factor as in the previous section:

$$P_{sc} = \frac{\alpha K \tau f_{CLK} \left(V_{DD} - 2V_T\right)^3}{12} . \tag{6.87}$$

The short-circuit power increases linearly with the switching frequency but increases (approximately) with the cube of the supply voltage. As discussed previously, it is possible to eliminate the crossover current and therefore the short-circuit power by increasing the absolute values of the threshold voltages so that $V_{DD} < \left(V_{TN} + \left|V_{TP}\right|\right)$.

6.9.3 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

$$I_D \approx \frac{(m-1)\mu_n \varepsilon_{ox} W}{t_{ox} L} \left(\frac{kT}{q} \right)^2 \exp\left(\frac{q(V_{GS} - V_T)}{mkT} \right), \tag{6.88}$$

where μ_n is electron mobility, ε_{OX} is permittivity of oxide, t_{OX} is oxide thickness, W is width of the MOSFET channel, L is length of the MOSFET channel, k is the Boltzmann constant, T is absolute temperature, q is electronic charge, and V_{GS} is gate-to-source bias voltage.

The unitless parameter m is given by

$$m = 1 + \frac{C_{dm}}{C_{ox}}, \tag{6.89}$$

where C_{dm} is the maximum depletion layer capacitance of the semiconductor under the oxide, and C_{OX} is the oxide capacitance. In typical MOSFETs, $1.5 < m < 2$, but for SOI MOSFETs, the value of m is close to 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 either logic state is equal to

$$I_{subthreshold} \approx K(m-1) \left(\frac{kT}{q} \right)^2 10^{-V_T/S}, \tag{6.90}$$

where the subthreshold swing S is given by

$$S \equiv \left(\frac{d(\log_{10} I_D)}{dV_{GS}} \right)^{-1} = \frac{mkT}{q} \ln(10) \tag{6.91}$$

and K is the device transconductance parameter of the symmetric n-channel and p-channel MOSFETs.

Therefore, the subthreshold power is approximately

$$P_{subthreshold} \approx V_{DD} K(m-1) \left(\frac{kT}{q} \right)^2 10^{-V_T/S}. \tag{6.92}$$

Typically, the absolute value of the threshold voltages is required to be three times the subthreshold swing 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. SOI MOSFETs have near ideal subthreshold characteristics and allow lower threshold voltages (~0.2 V or less).

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,

$$I_{pn} = I_S\left[\exp\left(\frac{qV}{nkT}\right)-1\right] \approx -I_S\,. \tag{6.93}$$

Each gate contributes a p-n junction leakage power equal to

$$P_{pn} \approx V_{DD} I_S\,. \tag{6.94}$$

The gate oxide leakage is attributable to quantum mechanical tunneling. Although a theoretical treatment of the oxide leakage current is beyond the scope of this book, this leakage component is negligible for CMOS circuits having gate oxide thicker than about 20 nm. The importance of this contribution is that it places a limit on the scaling of CMOS circuits using 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.

Example 6.13 Dissipation in a CMOS Inverter

Calculate and plot the dissipation versus the switching frequency for a symmetric CMOS gate with $V_{DD} = 2.5$ V, $V_T = 0.5$ V, and $K = 100\ \mu A/V^2$ and $C_L = 1$ pF.

Solution: The capacitance switching power is given by

$$P_{switch} = f\left(2.5V\right)^2\left(1\times 10^{-12}F\right) = f\left(6.2pJ\right)\cdot$$

Estimation of the short-circuit power is less straightforward because of its dependence on the rise and fall time for the input signal. A reasonable estimate may be obtained by assuming that the input signal has the same rise and fall time as the inverter under consideration, that is, $t_{RI} = t_R$ and $t_{FI} = t_F$. For this symmetric inverter, the rise and fall times are equal:

$$t_R = t_F = \frac{C_L}{K\left(V_{DD}-V_{TN}\right)}\left[\frac{2V_{TN}}{\left(V_{DD}-V_{TN}\right)}+\ln\left(\frac{19V_{DD}-20V_{TN}}{V_{DD}}\right)\right]$$

$$= \frac{1\times 10^{-12}F}{\left(100\times 10^{-6}A/V^2\right)\left(2.5V-0.5V\right)}\left[\frac{2\left(0.5V\right)}{\left(2.5V-0.5V\right)}+\ln\left(\frac{19\left(2.5V\right)-20\left(0.5V\right)}{2.5V}\right)\right]\cdot$$

$$= 16.1ns.$$

Therefore, using $t_{RI} = t_{FI} = \tau \approx 16.1ns$, we estimate the short-circuit power as

$$
\begin{aligned}
P_{sc} &= \frac{f\tau K\left(V_{DD} - 2V_T\right)^3}{12} \\
&= \frac{f\left(16.1\times10^{-9}ns\right)\left(100\mu A/V^2\right)\left(2.5V - 1.0V\right)^3}{12} \\
&= f\left(0.45pJ\right).
\end{aligned}
$$

Based on this estimate, the short-circuit power is less than 10% of the capacitance switching power.

The static dissipation is usually dominated by the subthreshold contribution. If it is assumed that m = 1.6, then

$$
\begin{aligned}
P_{subthreshold} &\approx V_{DD}K\left(m-1\right)\left(\frac{kT}{q}\right)^2 10^{-V_T/S} \\
&\approx \left(2.5V\right)\left(100\mu A/V^2\right)\left(1.6-1\right)\left(26mV\right)^2 10^{-5} \\
&= 1.0\times10^{-12}W.
\end{aligned}
$$

This contribution may usually be neglected except under standby conditions. (Typically, $I_S \sim 10^{-14}A$ so that $P_{pn} \sim 10fW$.)

For the example inverter, the dissipation as a function of the switching frequency characteristic may be calculated by

$$P = P_{subthreshold} + P_{switch} + P_{sc} = 6.8\times10^{-10}W + f\left(6.6\times10^{-12}J\right).$$

Considering only frequencies that allow the output to settle, the maximum switching frequency for these calculations is

$$f_{\max} \approx \frac{1}{2\max\left(t_R, t_F\right)} = \frac{1}{2\left(16.1ns\right)} = 31MHz.$$

Calculated results up to this frequency are provided in Figure 6.38.

6.10 Fan-Out

The maximum fan-out for CMOS circuits is determined entirely by dynamic considerations because the loading is primarily capacitive. The propagation delays increase with the number of load gates so that there is some maximum fan-out that corresponds to the longest allowable delays.

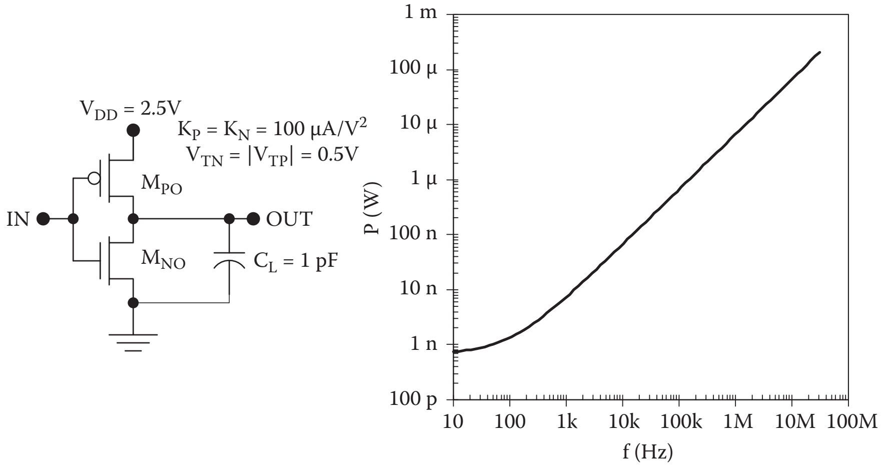


FIGURE 6.38
Calculated dissipation as a function of switching frequency for a CMOS inverter with $V_{DD} = 2.5$ V, $K_{NO} = K_{PO} = 100\ \mu A/V^2$, and $C_L = 1pF$, assuming $t_{RI} = t_{FI} = \tau \approx 16.1ns$.

Consider a symmetric CMOS inverter loaded by N similar CMOS gate circuits. Although the input capacitance per load gate is a function of the applied voltage, the worst case value may be estimated as

$$C_{in} = C_{gN} + C_{gP} = C_{ox}\left(W_N L_N + 2W_N L_{OV}\right) + C_{ox}\left(W_P L_P + 2W_P L_{OV}\right). \quad (6.95)$$

If the maximum allowable propagation delays are $t_{PLH,\max}$ and $t_{PHL,\max}$, then the maximum allowable load capacitance is

$$C_{L,\max} = \min\left(\frac{K_P\, t_{PLH,\max}}{\Gamma_P}, \frac{K_N\, t_{PHL,\max}}{\Gamma_N}\right). \quad (6.96)$$

The maximum fan-out is the largest integer satisfying

$$N_{MAX} \le \frac{C_{L,\max}}{C_{in}}, \quad (6.97)$$

or

$$N_{\max} = \min\left(\frac{K_P\, t_{PLH,\max}}{C_{in}\Gamma_P}, \frac{K_N\, t_{PHL,\max}}{C_{in}\Gamma_N}\right). \quad (6.98)$$

Typically, the maximum fan-out in a CMOS system is on the order of 10.

6.11 Circuit Delays as Functions of Fan-Out

For the estimation of the maximum fan-out in the previous section, we neglected the output capacitance C_{out} of the sending gate, and we also assumed that the load capacitance was entirely attributable to the receiving (load) gates. Often, however, the number of fan-out gates is small enough so that we should consider C_{out} as well as the parasitic capacitance of the interconnect wires $C_{\text{interconnect}}$. Then,

$$t_{PLH} = \frac{\left(C_{out} + NC_{in} + C_{\text{interconnect}}\right)\Gamma_P}{K_P} \tag{6.99}$$

and

$$t_{PHL} = \frac{\left(C_{out} + NC_{in} + C_{\text{interconnect}}\right)\Gamma_N}{K_N}. \tag{6.100}$$

In these expressions, the ratios Γ_P / K_P and Γ_N / K_N have units of Ω and may be considered to be effective switching resistances, allowing us to write the delay times in the practical forms

$$t_{PLH} = R_{swP}\left(C_{out} + NC_{in} + C_{\text{interconnect}}\right) \tag{6.101}$$

and

$$t_{PHL} = R_{swN}\left(C_{out} + NC_{in} + C_{\text{interconnect}}\right). \tag{6.102}$$

Here the effective switching resistances may be estimated from

$$R_{swP} = \frac{\Gamma_P}{K_P} = \frac{L_P}{W_P}\frac{t_{ox}}{\mu_p \varepsilon_{ox}}\frac{1}{\left(V_{DD} + V_{TP}\right)}\left[\frac{-2V_{TP}}{\left(V_{DD} + V_{TP}\right)} + \ln\left(\frac{3V_{DD} + 4V_{TP}}{V_{DD}}\right)\right] \tag{6.103}$$

and

$$R_{swN} = \frac{\Gamma_N}{K_N} = \frac{L_N}{W_N}\frac{t_{ox}}{\mu_n \varepsilon_{ox}}\frac{1}{\left(V_{DD} - V_{TN}\right)}\left[\frac{2V_{TN}}{\left(V_{DD} - V_{TN}\right)} + \ln\left(\frac{3V_{DD} - 4V_{TN}}{V_{DD}}\right)\right] \tag{6.104}$$

for long-channel MOS transistors. Generally, however, it is possible to determine the switching resistances empirically from measured propagation delays for various load capacitances:

$$R_{swP} = \frac{\partial t_{PLH}}{\partial C_L} \tag{6.105}$$

and

$$R_{swN} = \frac{\partial t_{PHL}}{\partial C_L}\ . \tag{6.106}$$

The *intrinsic delays* $t_{PLH,\mathrm{int}}$ and $t_{PHL,\mathrm{int}}$ are defined as the delays for a gate with unity fan-out in the limit of zero interconnect capacitance (meaning the sending and receiving gate are in close proximity). Thus,

$$t_{PLH,\mathrm{int}} = \frac{\left(C_{in} + C_{out}\right)\Gamma_P}{K_P} = R_{swP}\left(C_{in} + C_{out}\right) \tag{6.107}$$

and

$$t_{PHL,\mathrm{int}} = \frac{\left(C_{in} + C_{out}\right)\Gamma_N}{K_N} = R_{swN}\left(C_{in} + C_{out}\right). \tag{6.108}$$

Both C_{in} and C_{out} represent nonlinear voltage-dependent capacitances. However, we can estimate the worst-case input capacitance from the sum of the worst-case gate capacitances for the p-MOS and n-MOS devices:

$$C_{in} = C_{gN} + C_{gP} = C_{ox}\left(W_N L_N + 2W_N L_{OV}\right) + C_{ox}\left(W_P L_P + 2W_P L_{OV}\right). \tag{6.109}$$

The output capacitance consists mainly of the drain junction capacitances and drain-to-gate oxide capacitances. For the inverter,

$$C_{out} \approx \left(C_{dbN} + 2C_{gdN}\right) + \left(C_{dbP} + 2C_{gdP}\right), \tag{6.110}$$

where C_{dbN}, C_{dbP} are the drain-to-body capacitances and C_{gdN}, C_{gdP} are the gate-to-drain capacitances for the n-MOS and p-MOS devices, respectively. The factor of two accounts (approximately) for the Miller effect. For a typical CMOS inverter, C_{in} and C_{out} are comparable in value so that

$$C_{in} + C_{out} \approx 2C_{in}\ . \tag{6.111}$$

Example 6.14. Propagation Delays for Symmetric CMOS Inverter with N = 3

Estimate t_P for the symmetric CMOS inverter depicted in Figure 6.39 with three fan-out gates, neglecting the load capacitance of the interconnect. The fan-out gates are identical to the sending gate.

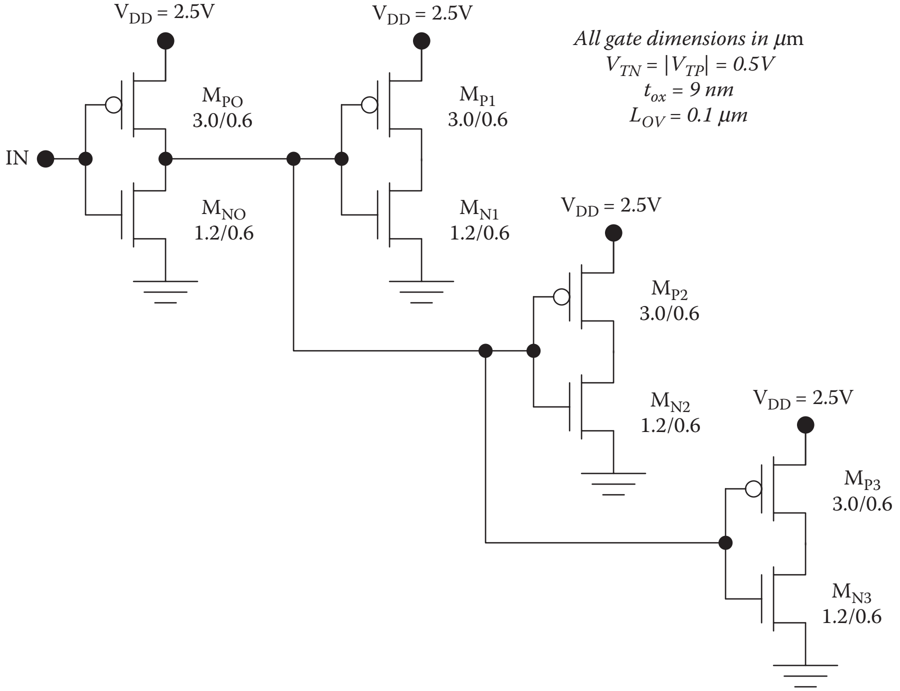


FIGURE 6.39
Example symmetric inverter with N = 3.

Solution: The effective switching resistances are

$$R_{swP} = \frac{L_P}{W_P}\frac{t_{ox}}{\mu_p\varepsilon_{ox}}\frac{1}{(V_{DD}+V_{TP})}\left[\frac{-2V_{TP}}{(V_{DD}+V_{TP})}+\ln\left(\frac{3V_{DD}+4V_{TP}}{V_{DD}}\right)\right]$$

$$= \frac{0.6\mu m}{3.0\mu m}\left(\frac{9\times10^{-7}cm}{(230cm^2V^{-1}s^{-1})(3.9)(8.85\times10^{-14}F)}\right)\left(\frac{1}{2.0V}\right)\left[\frac{1.0V}{(2.5V-0.5V)}+\ln\left(\frac{7.5V-2.0V}{2.5V}\right)\right]$$

$$= 1450\Omega$$

and

$$R_{swN} = \frac{L_N}{W_N}\frac{t_{ox}}{\mu_n\varepsilon_{ox}}\frac{1}{(V_{DD}+V_{TP})}\left[\frac{-2V_{TP}}{(V_{DD}+V_{TP})}+\ln\left(\frac{3V_{DD}+4V_{TP}}{V_{DD}}\right)\right]$$

$$= \frac{0.6\mu m}{1.2\mu m}\left(\frac{9\times10^{-7}cm}{(580cm^2V^{-1}s^{-1})(3.9)(8.85\times10^{-14}F)}\right)\left[\frac{1.0V}{(2.5V-0.5V)}+\ln\left(\frac{7.5V-2.0V}{2.5V}\right)\right]$$

$$= 1450\Omega$$

The oxide capacitance is $C_{ox} = \varepsilon_{ox} / t_{ox} = 3.83 \times 10^{-15} F / \mu m^2$, and the input capacitance per fan-out gate is

$$\begin{aligned} C_{in} &= C_{gN} + C_{gP} = C_{ox}\left(W_N L_N + 2W_N L_{OV}\right) + C_{ox}\left(W_P L_P + 2W_P L_{OV}\right)w \\ &= 3.83 \times 10^{-15} F / \mu m^2 \left[(1.2)(0.6) + 2(1.2)(0.1)\right] + \\ &\quad 3.83 \times 10^{-15} F / \mu m^2 \left[(3.0)(0.6) + 2(3.0)(0.1)\right] \\ &= 3.66 fF + 9.20 fF \approx 12.9 fF. \end{aligned}$$

If we make the approximations $C_{out} \approx C_{in}$ and $C_{\text{interconnect}} \approx 0$, then

$$\begin{aligned} t_{PLH} = t_{PHL} &= R_{swN}\left(C_{out} + NC_{in} + C_{\text{interconnect}}\right) \\ &= 1450\Omega\left(12.9 \times 10^{-15} F + 3\left(12.9 \times 10^{-15} F\right) + 0\right) = 75 ps. \end{aligned}$$

Example 6.15 Propagation Delays for a Minimum Size CMOS Inverter with N = 3

Estimate t_{PLH} and t_{PHL} for the minimum-size CMOS inverter with N = 3 as shown in Figure 6.40, neglecting the load capacitance of the interconnect. Assume that the fan-out gates are identical to the sending gate.

Solution: The effective switching resistances are

$$\begin{aligned} R_{swP} &= \frac{L_P}{W_P} \frac{t_{ox}}{\mu_p \varepsilon_{ox}} \frac{1}{\left(V_{DD} + V_{TP}\right)} \left[\frac{-2V_{TP}}{\left(V_{DD} + V_{TP}\right)} + \ln\left(\frac{3V_{DD} + 4V_{TP}}{V_{DD}}\right)\right] \\ &= \frac{0.6\mu m}{1.2\mu m}\left(\frac{9 \times 10^{-7} cm}{\left(230 cm^2 V^{-1} s^{-1}\right)(3.9)\left(8.85 \times 10^{-14} F\right)}\right)\left(\frac{1}{2.0V}\right) \\ &\qquad \left[\frac{1.0V}{\left(2.5V - 0.5V\right)} + \ln\left(\frac{7.5V - 2.0V}{2.5V}\right)\right] \\ &= 3650\Omega \end{aligned}$$

and

$$\begin{aligned} R_{swN} &= \frac{L_N}{W_N} \frac{t_{ox}}{\mu_n \varepsilon_{ox}} \frac{1}{\left(V_{DD} + V_{TP}\right)} \left[\frac{-2V_{TP}}{\left(V_{DD} + V_{TP}\right)} + \ln\left(\frac{3V_{DD} + 4V_{TP}}{V_{DD}}\right)\right] \\ &= \frac{0.6\mu m}{1.2\mu m}\left(\frac{9 \times 10^{-7} cm}{\left(580 cm^2 V^{-1} s^{-1}\right)(3.9)\left(8.85 \times 10^{-14} F\right)}\right)\left[\frac{1.0V}{\left(2.5V - 0.5V\right)} + \ln\left(\frac{7.5V - 2.0V}{2.5V}\right)\right] \\ &= 1450\Omega. \end{aligned}$$

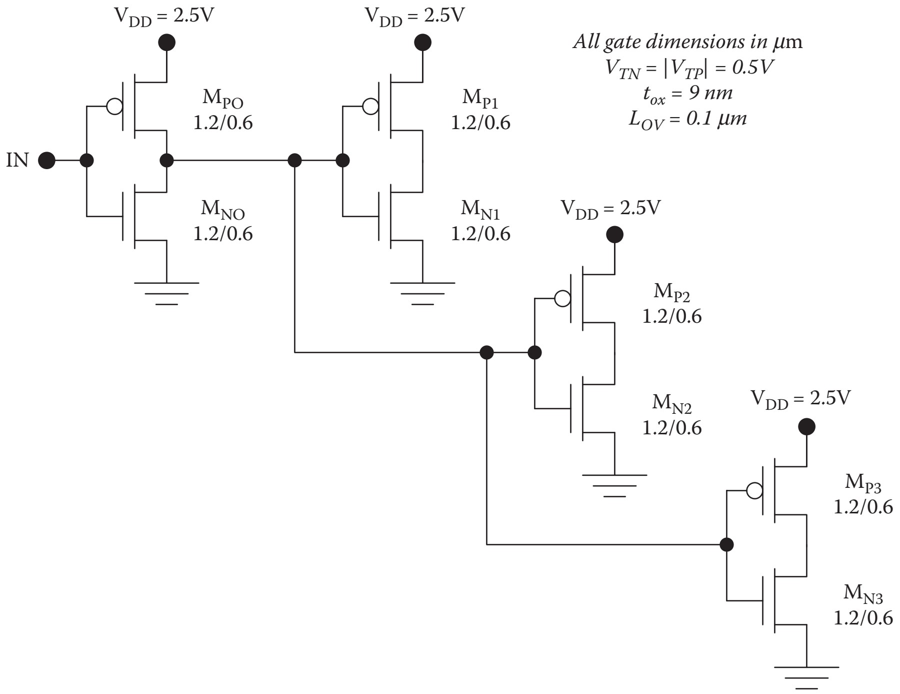


FIGURE 6.40
Example minimum-size inverter with N = 3.

In the minimum-size inverter circuit, the p-MOS device has ~2.5 times the effective switching resistance of the n-MOS device. The oxide capacitance is $C_{ox} = \varepsilon_{ox} / t_{ox} = 3.83 \times 10^{-15} F / \mu m^2$, and the input capacitance per fan-out gate is

$$\begin{aligned}
C_{in} &= C_{gN} + C_{gP} = C_{ox}\left(W_N L_N + 2W_N L_{OV}\right) + C_{ox}\left(W_P L_P + 2W_P L_{OV}\right) \\
&= 3.83 \times 10^{-15} F / \mu m^2 \left[(1.2)(0.6) + 2(1.2)(0.1)\right] + \\
&\quad 3.83 \times 10^{-15} F / \mu m^2 \left[(1.2)(0.6) + 2(1.2)(0.1)\right] \\
&= 3.7fF + 3.7fF \approx 7.4fF.
\end{aligned}$$

Therefore, because the p-MOS device has been minimum size, the input capacitance is reduced by about 40% compared with the symmetric inverter. If we make the approximations $C_{out} \approx C_{in}$ and $C_{\text{interconnect}} \approx 0$, then

$$\begin{aligned}
t_{PLH} &= R_{swP}\left(C_{out} + NC_{in} + C_{\text{interconnect}}\right) \\
&= 3650\Omega\left(7.4 \times 10^{-15} F + 3\left(7.4 \times 10^{-15} F\right) + 0\right) = 108ps
\end{aligned}$$

and

$$t_{PHL} = R_{swN}\left(C_{out} + NC_{in} + C_{\text{interconnect}}\right)$$
$$= 1450\Omega\left(7.4\times10^{-15}F + 3\left(7.4\times10^{-15}F\right) + 0\right) = 43ps.$$

Compared with the symmetric inverter, the minimum-size circuit has an increased t_{PLH} attributable to the weaker p-MOS device. However, the reduced input capacitance for the load gates compensates somewhat for the degraded p-MOS current drive so t_{PLH} is degraded only by about 35%. At the same time, t_{PHL} is actually improved by the reduction in the load capacitance. The sum t_{PHL} + t_{PLH} is approximately the same as for the symmetric inverter, but this is achieved with less circuit area.

6.12 CMOS Ring Oscillator

In the CMOS ring oscillator, each gate experiences unity fan-out and so the frequency of oscillation may be determined from the intrinsic delay times. For a CMOS ring oscillator with N stages,

$$f \approx \frac{1}{N\left(t_{PLH,\text{int}} + t_{PHL,\text{int}}\right)}. \tag{6.112}$$

A more refined estimate may be obtained by accounting for the finite rise and fall time at the input to each stage, which will typically decrease the frequency of oscillation by 20%.

6.13 CMOS Inverter Design

Figures 6.41 and 6.42 show example layout designs for minimum-size and symmetric CMOS inverters, respectively. In each case, the p-MOS transistor is fabricated in an n-well, whereas the n-MOS transistor is fabricated directly in the p-type substrate. Using scalable design rules with a minimum gate length of 2X, the minimum surround of the p-MOS active region by the n-well is 5X, and the minimum separation between the n-well and the active region for the n-MOS device is 5X. The n-well must have an electrical connection to V_{DD} to prevent forward bias of the n-well/p-substrate junction. For the same reason, the p-type substrate be tied to ground.

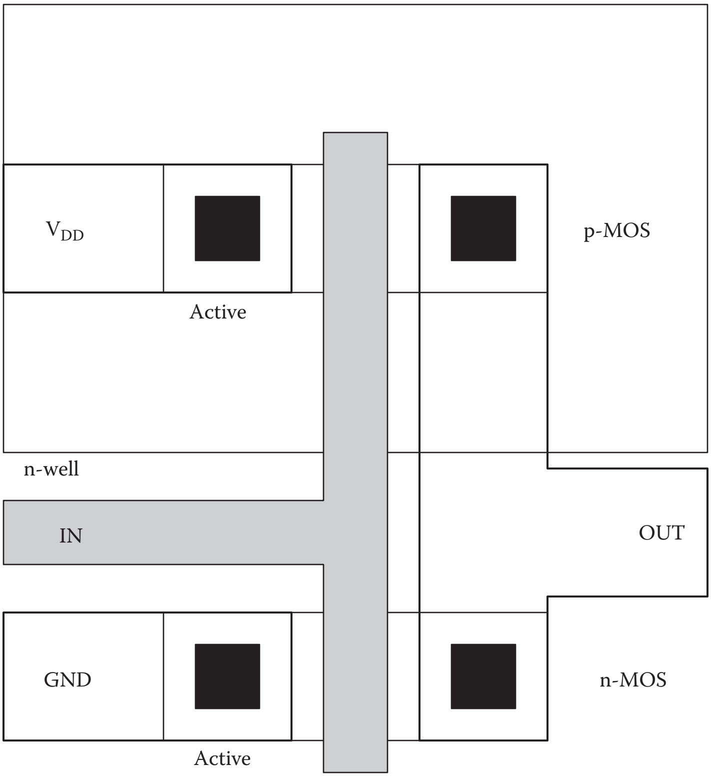


FIGURE 6.41
Layout of minimum-size CMOS inverter.

6.14 CMOS NAND Circuits

Realization of the NAND function in CMOS requires the series connection of n-MOSFETs in the pull-down branch and parallel connection of p-MOSFETs in the pull-up branch. The two-way and three-way NAND circuits are shown in Figures 6.43 and 6.44, respectively, and it can be seen that a NAND gate with a fan-in of M requires 2M transistors (one n-MOSFET and one p-MOSFET per input). The output of the NAND circuit will go low (to ~0 V) only if all of the n-MOSFETs are on and all of the p-MOSFETs are off; this only occurs with logic one applied to all inputs. If a single input is brought to zero, the associated n-MOSFET will be cutoff whereas the associated p-MOSFET will be linear, thus bringing the output to V_{DD}.

6.14.1 Sizing of Transistors in a CMOS NAND Gate

In a NAND gate with M inputs, one approach to sizing is to size up the n-MOS transistors by a factor of approximately M to maintain static and

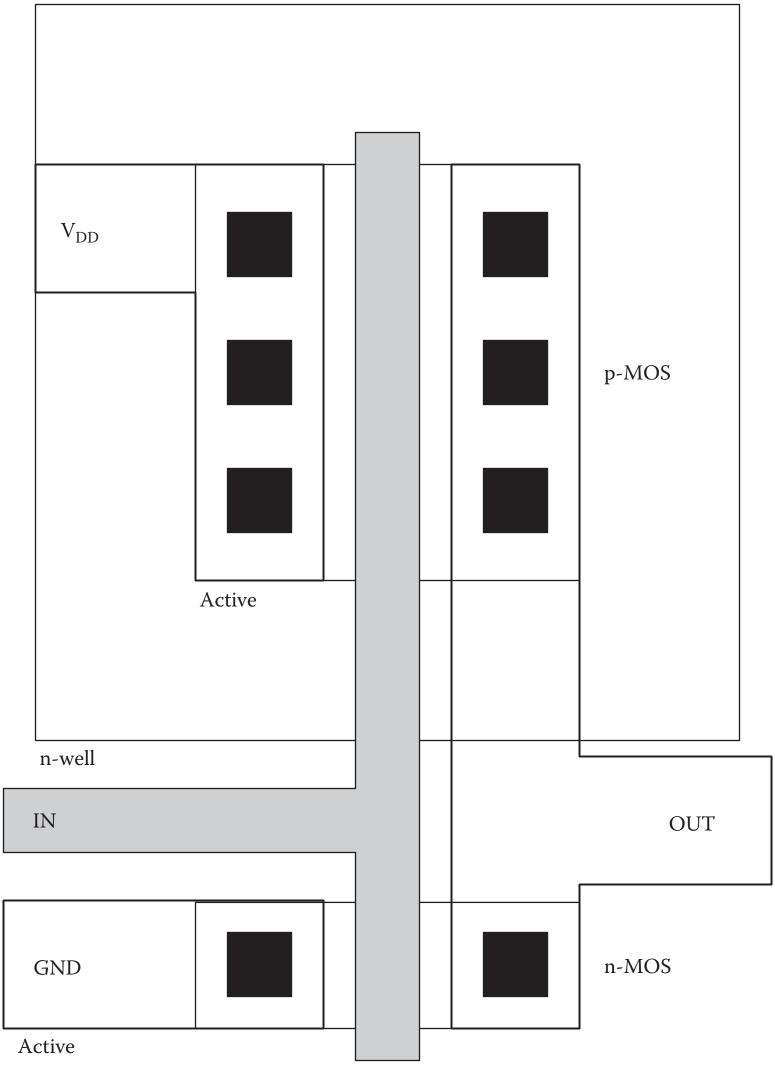


FIGURE 6.42
Layout design of symmetric CMOS inverter.

dynamic performance characteristics comparable with the reference inverter circuit made in the same technology (see Figure 6.45). This is motivated by the fact that the M series-connected n-MOS devices behave approximately as a single transistor with M times the gate length, and they therefore need M times the width to maintain the same current drive capability.

This can also be understood from the point of view that all of the n-MOS devices must conduct in series for the output to go low. The drain resistances of these (linear) transistors add together. Therefore, to maintain the same

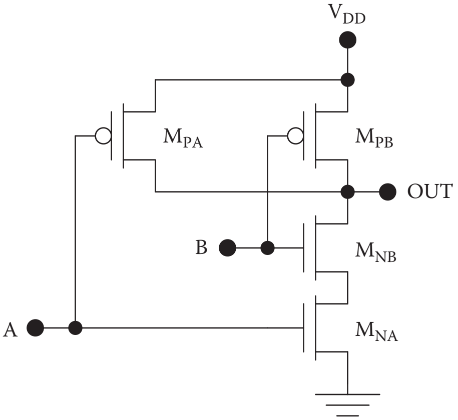


FIGURE 6.43
Two-way CMOS NAND gate.

total "on" resistance R_{on} as in the inverter, each of the M series transistors must contribute a resistance of R_{on}/M, and this requires that their widths be scaled up by a factor of M.

Following the same line of reasoning, the p-MOS transistors need not be scaled up in the NAND gate compared with the reference inverter. One or more p-MOSFETs conduct when the output goes high, but the simultaneous conduction of two or more p-MOSFETs will only improve the pull-up current capability.

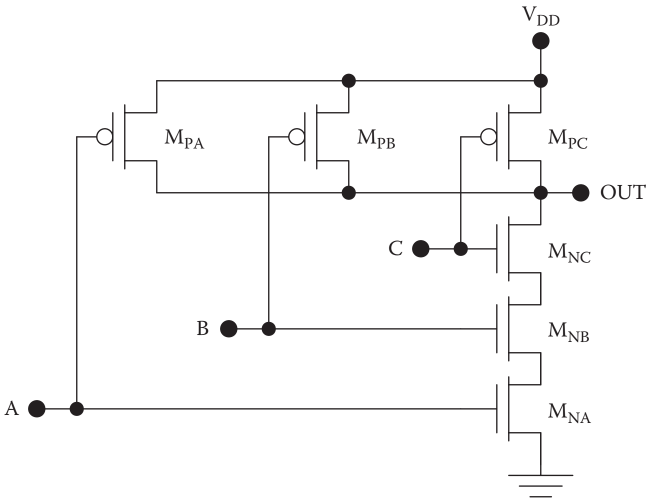


FIGURE 6.44
Three-way CMOS NAND circuit.

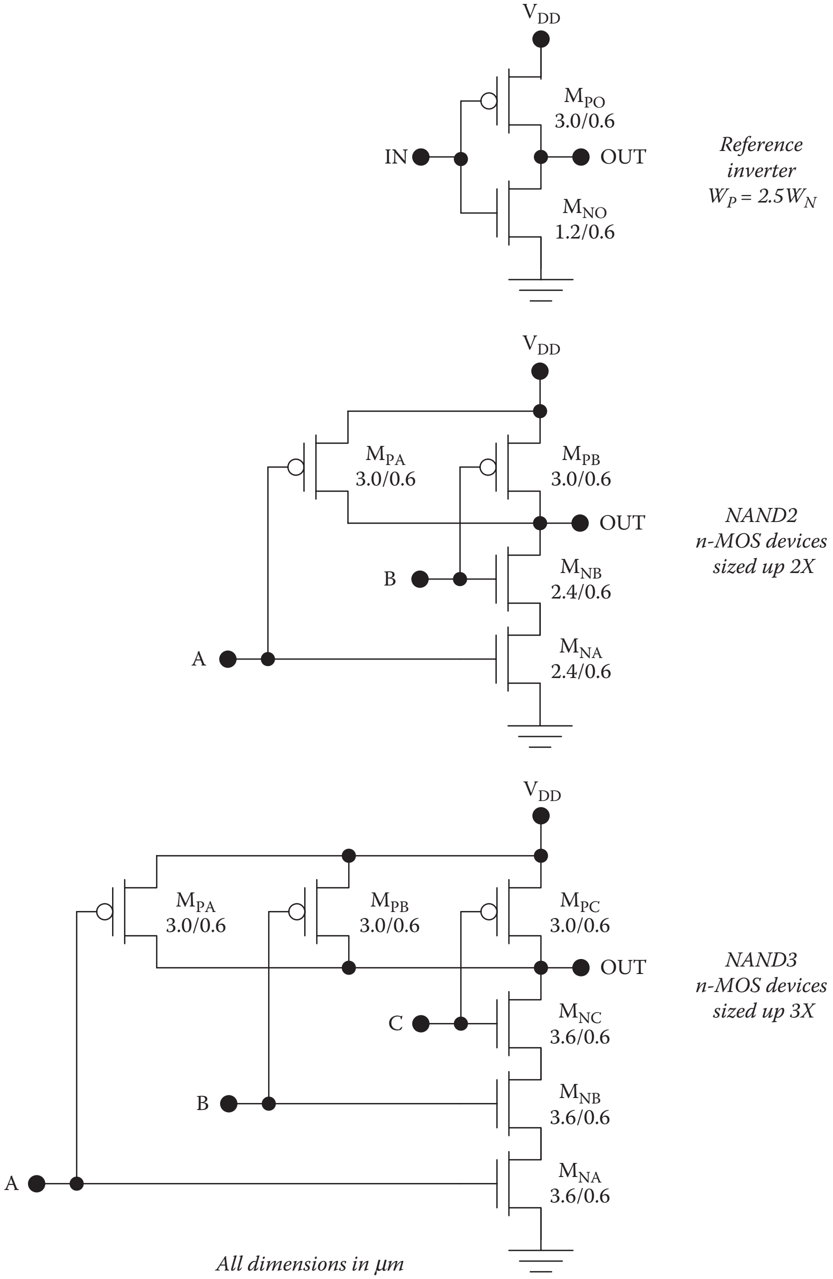


FIGURE 6.45
Sizing of transistors in CMOS NAND gates. In the symmetric reference inverter, $W_N / W_P = 1/2.5$ to compensate for the lower mobility of holes compared with electrons. In the M-way NAND gate, the n-MOS transistors are sized up by a factor of M compared with the reference inverter. To achieve symmetric performance, $W_N / W_P = M/2.5$.

6.14.2 Static Characteristics of the CMOS NAND Gate

There is no unique voltage transfer characteristic for the CMOS NAND gate; instead, the M-way NAND gate has M cases of the voltage transfer characteristic based on the number of inputs that is varied. This is illustrated in Figure 6.46, which shows voltage transfer characteristics for a three-way CMOS NAND circuit having $K_P = 100\mu A / V^2$, $K_N = 3K_P = 300\mu A / V^2$, $V_{TN} = |V_{TP}| = 0.5V$, $\gamma_P = \gamma_N = 0.1V^{1/2}$, and $\lambda_P = \lambda_N = 0.01V^{-1}$. In case I, one of the inputs is varied, whereas the other two are fixed at V_{DD}. Because two of the three p-MOS devices stay off, the characteristic is similar to that for an inverter in which $K_R = K_N/K_P = 3$, and the switching threshold is 1.14 V. In case II, two of the inputs are varied, whereas the other is fixed at V_{DD}. Because

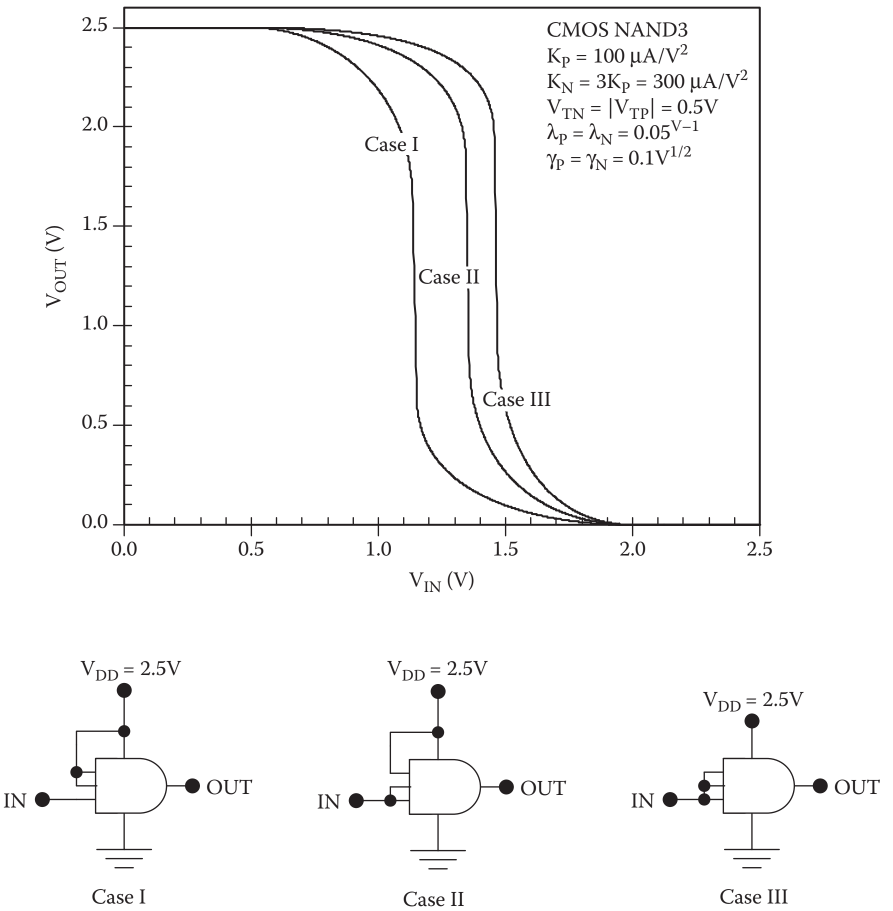


FIGURE 6.46
Static characteristics for a three-way NAND gate with $K_N = 3K_P$.

two p-MOS devices are conducting, the crossover current is increased at any particular value of input voltage and so the characteristic shifts to the right with a switching threshold of 1.35 V. In case III, all three inputs are varied, causing the characteristic to shift further to the right and the switching threshold is 1.46 V.

The three cases described above are based entirely on how many inputs are varied; there are subcases that differ based on which of the inputs are varied because the three n-MOS devices behave differently. This arises because the sources of the three n-MOS transistors are at different electric potentials, as can be seen in the circuit diagram of Figure 6.47. Notice that only the bottom n-MOS device in the pull-down stack has its source grounded, whereas all of the n-MOS device bodies are tied to ground. The bottom n-MOS transistor M_{NA} has zero body-source bias and so there is no body effect. The source of transistor M_{NB} is at a potential equal to the drain-to-source voltage for the transistor below it, so $V_{BSNB} = -V_{DSNA}$ and the body effect will increase the threshold voltage for this device. There is also a body effect for the top n-MOS device M_{NC} because $V_{BSNC} = -(V_{DSNA} + V_{DSNB})$.

Another effect associated with the series connection of the n-MOS transistors is that the gate-to-source voltages of the upper devices are decreased because of their nonzero source voltages. For the bottom n-MOS device, the gate-to-source voltage is equal to the associated input voltage $V_{GSNA} = V_{INA}$, but for the middle transistor $V_{GSNB} = V_{INB} - V_{DSNA}$ and for the top transistor $V_{GSNC} = V_{INC} - (V_{DSNA} + V_{DSNB})$.

Because the upper n-MOS transistors in the NAND gate have increased threshold voltages (as a consequence of the body effect) and reduced gate-to-source

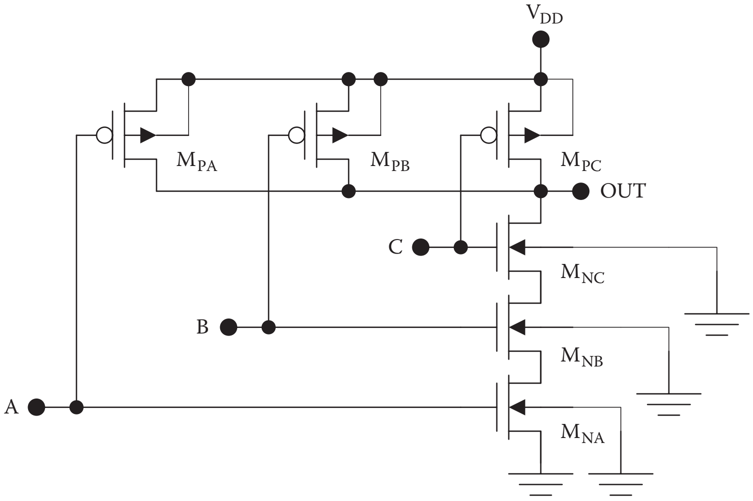


FIGURE 6.47
Three-way NAND gate showing transistor body connections.

voltages, there are three subcases for switching a single input. We will refer to these as subcases Ia, Ib, and Ic, for which input A, B, or C is varied, respectively. Similarly, there are subcases IIa, IIb, and IIc, for which input A, B, or C is held constant at V_{DD}. On the other hand, all inputs are tied together in case III so there are no subcases.

For higher-order NAND gates, the situation is complicated considerably. For example, the four-way CMOS NAND gate has 15 distinct voltage transfer characteristics. This makes it necessary to consider the most extreme cases (with the lowest and highest switching thresholds) to determine the worst-case noise margins. The noise margins so determined will generally be inferior to those of the reference inverter, even if the transistors in the NAND gate are scaled appropriately.

6.14.3 Dynamic Characteristics of the CMOS NAND Gate

The transient response and low-to-high propagation delay for a CMOS NAND gate depends strongly on how many of the inputs make voltage transitions as illustrated in Figure 6.48. This figure shows transient characteristics for a three-way NAND circuit in which $K_P = 100\mu A / V^2$, $K_N = 3K_P = 300\mu A / V^2$, $V_{TN} = |V_{TP}| = 0.5V$, $\gamma_P = \gamma_N = 0.1V^{1/2}$, and $\lambda_P = \lambda_N = 0.01V^{-1}$. In case I, one of the inputs is varied, whereas the other two are fixed at V_{DD}. Because two of the three p-MOS devices stay off, this results in the worst-case low-to-high propagation delay, which is approximately equal to that of the reference inverter $\left(t_{PLH} = 6.3ns, \, case\ I\right)$. In case II, two of the inputs are varied, whereas the other is fixed at V_{DD}. With two p-MOSFETs conducting, the low-to-high propagation delay is approximately half of the value for case I $\left(t_{PLH} = 3.4ns, \, case\ II\right)$. Finally, in case III, all three inputs are switched, and, with three p-MOS devices conducting, the propagation delay is approximately one-third of the worst case value for case I $\left(t_{PLH} = 2.4ns; \, case\ III\right)$. However, the high-to-low propagation delay is nearly the same for all three cases $\left(t_{PHL} = 6.7ns; \, cases\ I,\ II,\ and\ III\right)$. The subtle differences between the t_{PHL} values arise because of unequal body effects and gate-to-source voltages for the stacked n-MOS devices.

6.15 CMOS NOR Circuits

A CMOS NOR gate is realized by placing the p-MOS transistors in series and the n-MOS transistors in parallel as shown in Figure 6.49 (two-way circuit). If any input goes high, the associated n-MOS device will be linear, whereas the associated p-MOS device will be cutoff, so the output will go low. One approach to scaling of the p-MOS transistors is to size them up by a factor

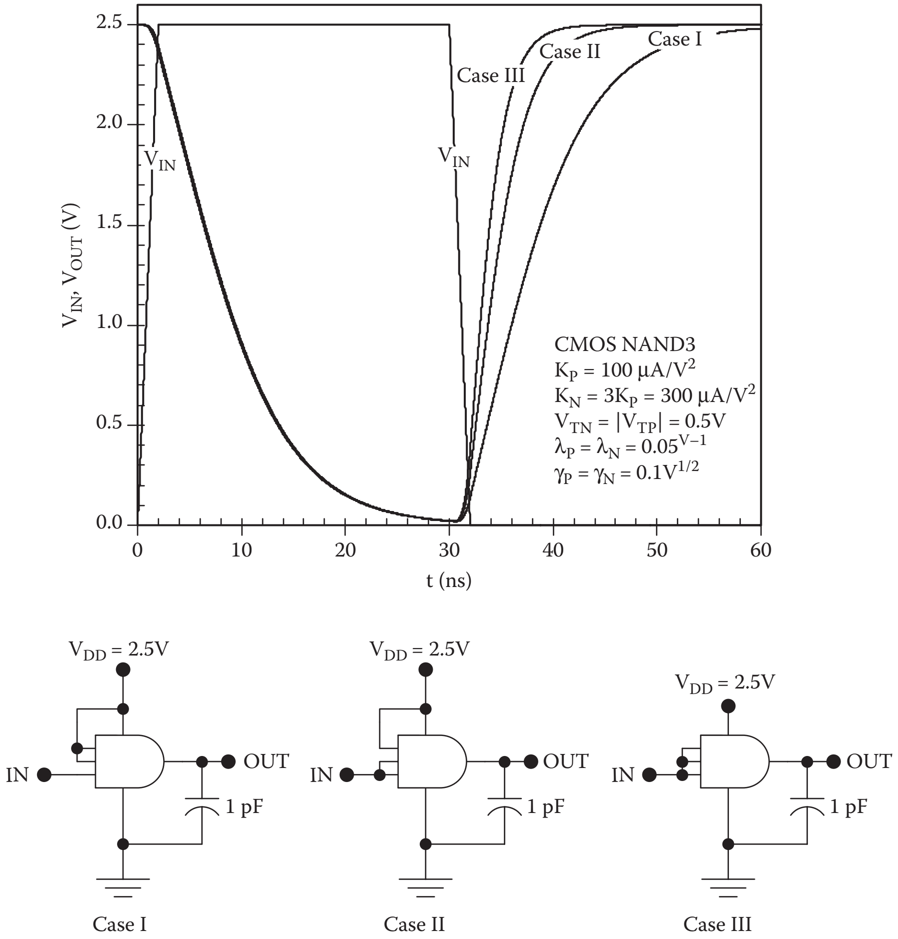


FIGURE 6.48
Transient characteristics for a three-way NAND gate with $K_N = 3K_P$ and a lumped 1 pF load.

equal to the fan-in M. (This is illustrated in Figure 6.50.) If symmetric characteristics are desired, an NOR gate therefore takes up more chip area than a NAND gate with the same fan-in. This is because, in the reference symmetric inverter, the p-MOS devices are already wider than the n-MOS devices by a factor of ~2.5 to compensate for the lower mobility of holes compared with electrons. Therefore, scaling up the p-MOSFETS (in a NOR gate) adds more total gate width than scaling up the n-MOSFETs (in a NAND gate).

The M-way NOR circuit exhibits M cases of the static voltage transfer characteristic, depending on how many of the inputs are varied. As with the NAND, there are subcases for the voltage transfer characteristic depending on which of the inputs are varied because the sources of the p-MOS devices

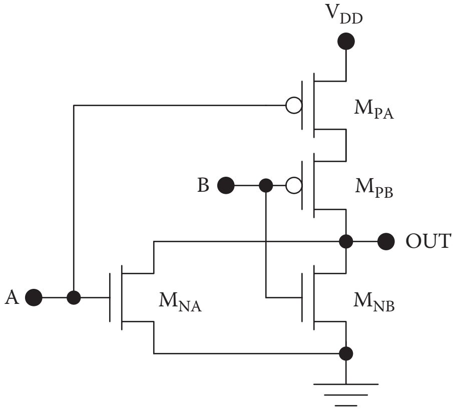


FIGURE 6.49
Two-way CMOS NOR gate.

are at different electric potentials, but all of the p-MOS devices have their bodies tied to V_{DD}. All but the top p-MOS device will have their threshold voltages shifted to be more negative by the body effect. Also, all but the top p-MOS device will experience less negative gate-to-source voltages. These effects will act together such that the worst-case noise margins will be worse than for the reference inverter, even with appropriate sizing of the p-MOS devices.

The switching speed performance of the NOR gate will also depend on how many of the inputs are switched, and t_{PHL} will vary greatly in these three cases. On the other hand, t_{PLH} will be substantially the same in the three cases. However, appropriate sizing of the p-MOS transistors (scaling their widths by a factor equal to the fan-in) will render t_{PLH} and the worst case of t_{PHL} equal to the propagation delays of the reference inverter circuit.

6.16 Other Logic Functions in CMOS

Complex logic functions can be implemented in CMOS by the combination of parallel and series branches of n- and p-MOSFETs. An example is shown in Figure 6.51.

The logic function performed by this gate is

$$Y = \overline{AB + CD}\,, \tag{6.113}$$

and the gate level representation of this circuit is as shown in Figure 6.52.

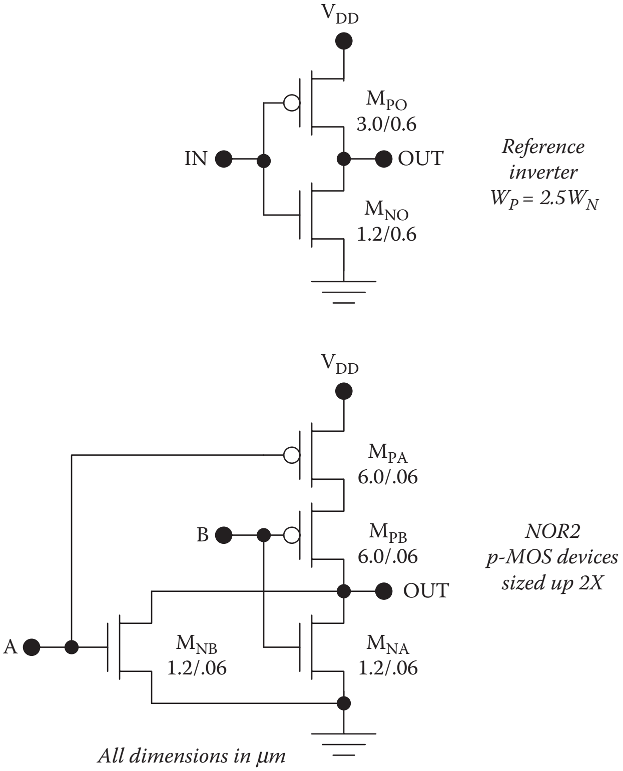


FIGURE 6.50
Sizing of transistors in CMOS NOR gates. Compared with the reference inverter circuit, the p-MOS devices must be sized up by a factor equal to the fan-in M.

In CMOS, the XOR function is implemented using an AND-OR-INVERT approach. The realization is shown in Figure 6.53, and the gate-level representation appears in Figure 6.54. This circuit performs the function

$$Y = \overline{AB + \overline{A}\,\overline{B}} = A \oplus B\,. \tag{6.114}$$

Implementation of the XOR2 function in CMOS is inefficient, requiring 12 MOSFETs (compared with five MOSFETs in the NMOS realization).

Other more complex logic functions may be implemented in CMOS by the extension of the AND-OR-INVERT concept. This requires appropriate scaling of the n-MOS and p-MOS transistors to maintain characteristics similar to the reference inverter. However, both the static and dynamic performance estimates are complicated by body effects and variable voltages at the transistor sources, so that worst-case analyses should always be applied.

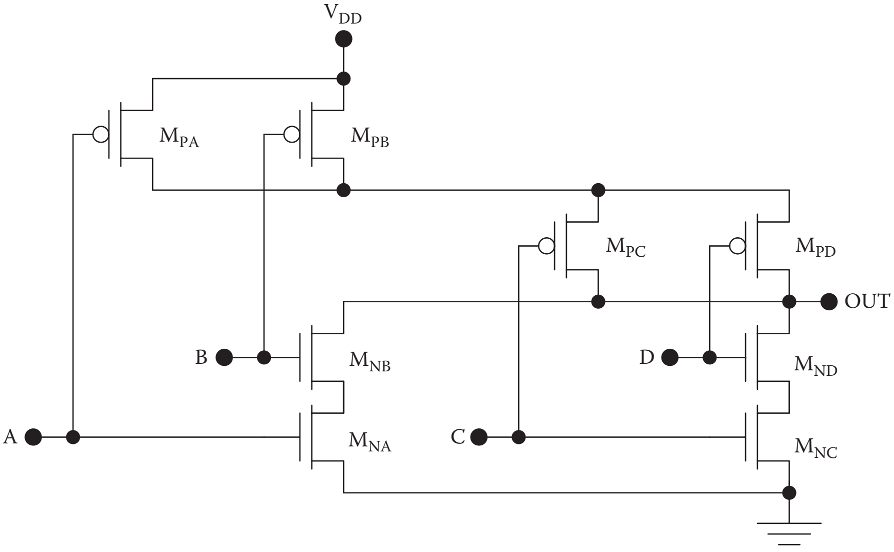


FIGURE 6.51
CMOS AND-OR-INVERT gate.

6.16.1 Transistor Sizing in CMOS AND-OR-INVERT Gates

Normally, all transistors on a wafer will have identical gate lengths that are imposed by the minimum feature size for the fabrication technology. However, the transistor widths are adjusted for the desired current drive capability. In the case of a general AND-OR-INVERT circuit, the transistor widths may be related to those in a reference inverter that has the desired electrical characteristics (switching threshold, propagation delays, rise and fall times). Suppose the transistor widths in the reference inverter circuit are W_{PR} and W_{NR} for the p-MOS and n-MOS transistors, respectively. In the general AND-OR-INVERT circuit, the ith p-MOS device should be sized with a width given by $M_{Pi}W_{PR}$, where M_{Pi} is the maximum number of p-channel transistors between the output node and V_{DD}, for any path including the ith p-MOS transistor. In like manner, the jth n-MOS device should be sized with

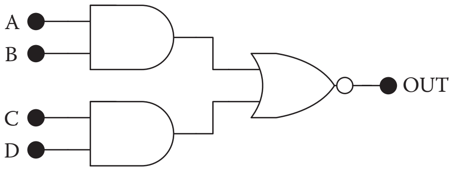


FIGURE 6.52
Gate-level representation of the CMOS AND-OR-INVERT circuit shown in Figure 6.51.

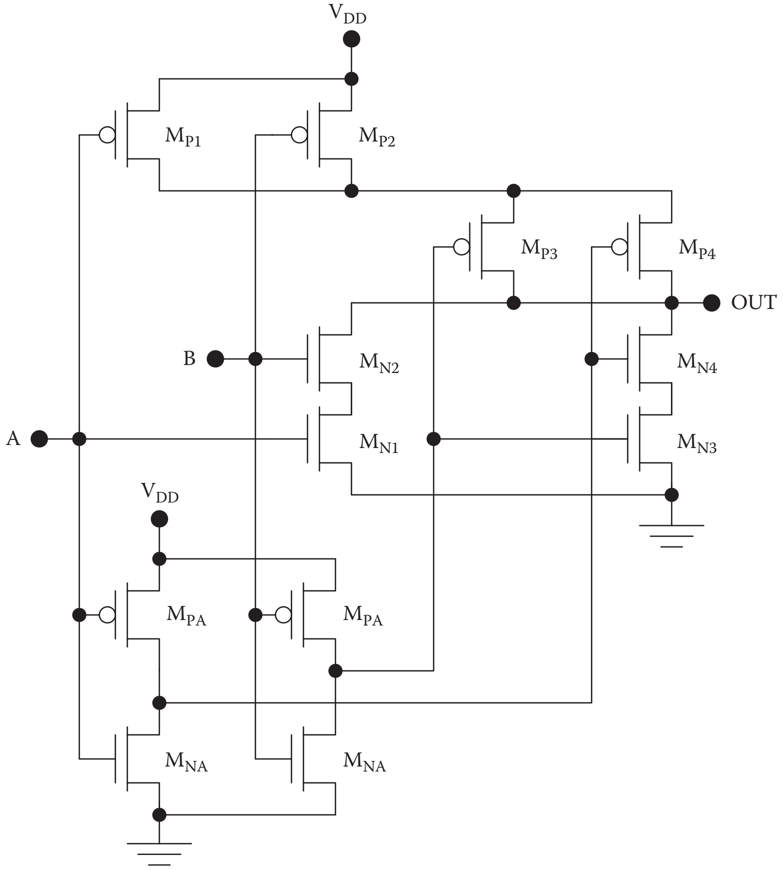


FIGURE 6.53
CMOS XOR circuit.

a width given by $M_{Nj}W_{NR}$, where M_{Nj} is the maximum number of n-channel transistors between the output node and ground, for any path including the jth n-MOS device. It is important to note that the scaling factors may not be the same for all n-channel transistors or for all p-channel transistors.

6.17 74HC Series CMOS

The 74HC series of CMOS is a common family of SSI to MSI logic components. These circuits are *double buffered* and use MOS devices with polysilicon gates, 3 μm gate lengths, and 60 nm thick gate oxide.

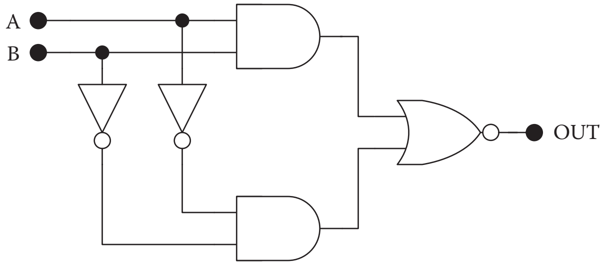


FIGURE 6.54
Gate-level representation of the CMOS XOR circuit.

The 74HC00 quad two-input NAND gate comprises four identical circuits like the one shown in Figure 6.55. This circuit is double buffered by two inverters, which do not alter the overall logic function but greatly improve the voltage transfer characteristic.

The basic characteristics of 74HC high-speed CMOS gates are summarized in Table 6.3. The maximum supply voltage of 5.5 V is limited by the MOSFET breakdown characteristics. The typical propagation delay of 10 ns with $C_L = 15$ pF corresponds to an effective switching resistance of $R_{sw} = 670\Omega$.

Figure 6.56 shows how double buffering sharpens the voltage transfer characteristic, rendering it nearly ideal with a sharp transition at $V_{DD}/2$. This diagram shows the voltage transfer characteristic derived from the

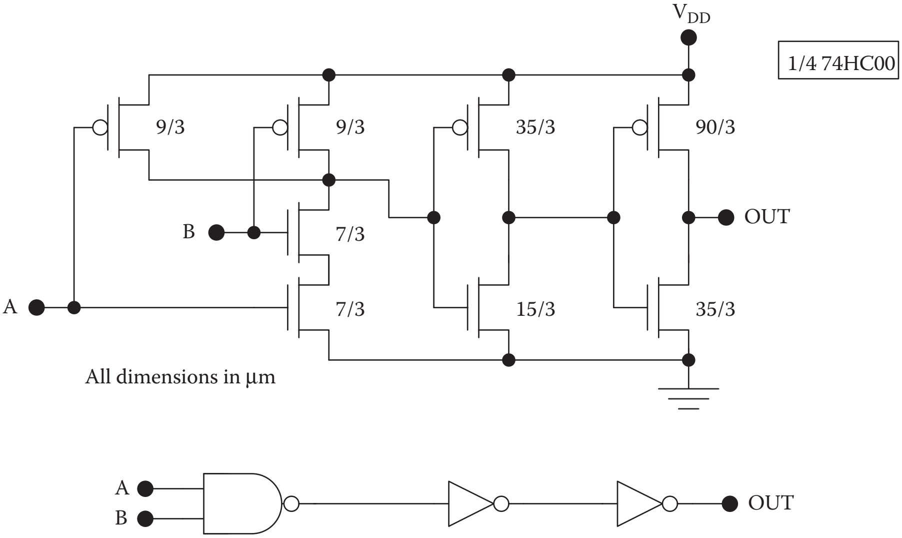


FIGURE 6.55
74HC Series NAND2 circuit (¼ of the 74HC00 quad two-input NAND gate).

TABLE 6.3

Characteristics of the 74HC Circuit Family

74HC series CMOS	
Gate material	Polysilicon
Gate length	3 μm
Oxide thickness	60 nm
Supply voltage	4.5–5.5 V
Propagation delay (C_L = 15 pF)	10 ns

outputs of stage one, stage two, and stage three (the output stage) for a $\frac{1}{6}$ 74HC04 double-buffered inverter. All three characteristics exhibit switching thresholds very close to $V_{DD}/2 = 2.5V$; however, the slopes of the characteristic at the switching threshold $\partial V_{OUT}/\partial V_{IN}$ are –24, 560, and –8600 for OUT1, OUT2, and OUT3, respectively. More importantly, however, the double buffering moves V_{IL} and V_{IH} closer together, thus improving the noise margins. The critical voltages V_{IL}, V_{IH} are 1.96 V, 3.08 V for OUT1, 2.39 V, 2.65 V for OUT2, and 2.50 V, 2.53 V for OUT3. In VLSI circuits, buffering can provide improved dynamic response as well as sharper static characteristics.

Example 6.16 Propagation Delay for a 74HC Inverter with C_L = 15 pF

Estimate the propagation delay for the 74HC CMOS inverter ($\frac{1}{6}$ 74HC04) as shown in Figure 6.57 with a 15 pF load. Assume that the gate-drain and drain-source overlaps are 0.2 μm for all transistors.

Solution: The process transconductance parameters for the p-MOSFETs and n-MOSFETs are

$$k_P^{'} = \frac{\mu_p \varepsilon_{ox}}{t_{ox}} = \frac{\left(230cm^2/Vs\right)(3.9)\left(8.85\times10^{-14}F/cm\right)}{60\times10^{-7}cm} = 13.4\mu A/V^2$$

and

$$k_N^{'} = \frac{\mu_n \varepsilon_{ox}}{t_{ox}} = \frac{\left(580cm^2/Vs\right)(3.9)\left(8.85\times10^{-14}F/cm\right)}{60\times10^{-7}cm} = 33.4\mu A/V^2,$$

respectively, and the oxide capacitance per unit area is

$$C_{ox} = \frac{\varepsilon_{ox}}{t_{ox}} = \frac{3.9\left(8.85\times10^{-14}F/cm\right)}{60\times10^{-7}cm} = 5.75\times10^{-8}F/cm^2 = 0.575fF/\mu m^2.$$

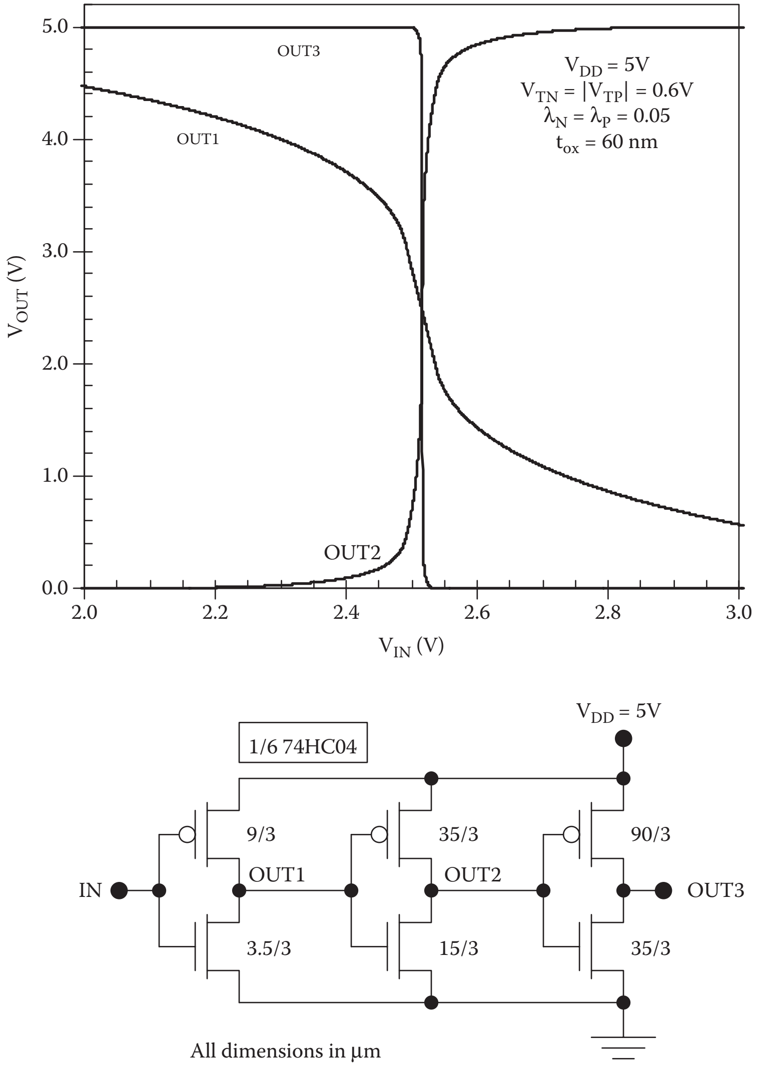


FIGURE 6.56
Voltage transfer characteristics for a 74HC series double-buffered inverter ($\frac{1}{6}$ 74HC04).

The delay factors are equal because of the symmetry in the threshold voltages:

$$\Gamma_P = \Gamma_N = \frac{1}{(V_{DD} - V_{TN})}\left[\frac{2V_{TN}}{(V_{DD} - V_{TN})} + \ln\left(\frac{3V_{DD} - 4V_{TN}}{V_{DD}}\right)\right]$$

$$= \frac{1}{(5V - 0.6V)}\left[\frac{2(0.6V)}{(5V - 0.6V)} + \ln\left(\frac{3(5V) - 4(0.6V)}{5V}\right)\right] = 0.272V^{-1}.$$

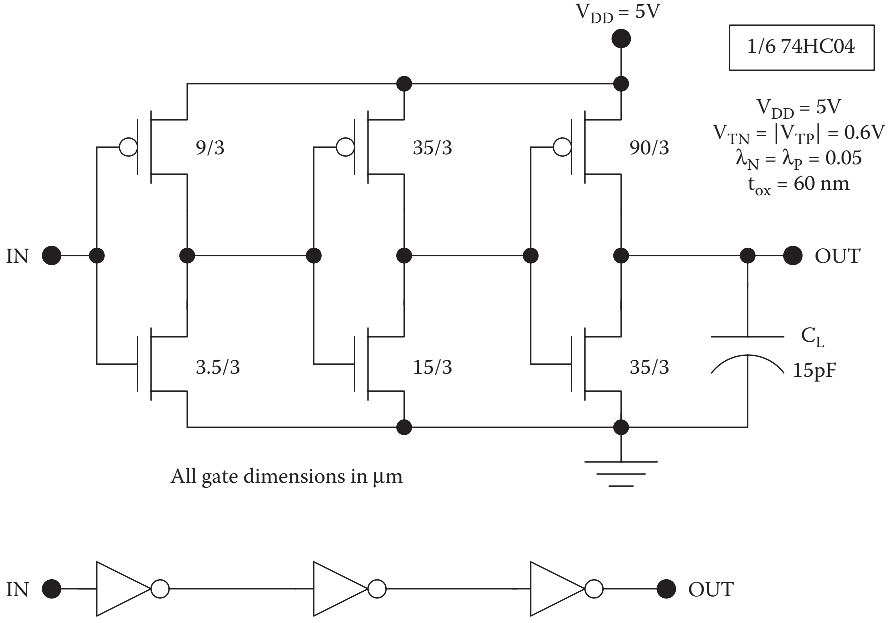


FIGURE 6.57
74HC inverter ($\frac{1}{6}$ 74HC04) with 15 pF load.

For the first stage, the device transconductance parameters are

$$K_{P1} = k_P^{'} \frac{W_{P1}}{L_{P1}} = 13.4\mu A / V^2 \left(\frac{9}{3}\right) = 40\mu A / V^2$$

and

$$K_{N1} = k_N^{'} \frac{W_{N1}}{L_{N1}} = 33.4\mu A / V^2 \left(\frac{3.5}{3}\right) = 40\mu A / V^2 ,$$

respectively, and the load capacitance seen by the first stage is the input capacitance for the second stage,

$$\begin{aligned} C_{L1} &= C_{gP2} + C_{gN2} = C_{ox}\left(W_{P2}L_{P2} + 2W_{P2}L_{OV}\right) + C_{ox}\left(W_{N2}L_{N2} + 2W_{N2}L_{OV}\right) \\ &= 0.575fF\left[(35)(3) + 2(35)(0.2)\right] + 0.575fF\left[(15)(3) + 2(15)(0.2)\right] = 98fF. \end{aligned}$$

The propagation delay for the first stage is therefore

$$t_{P1} = \frac{C_{L1}\Gamma}{K_1} = \frac{\left(98\times 10^{-15}F\right)\left(0.272V^{-1}\right)}{40\times 10^{-6} A / V^2} = 0.67ns.$$

Similarly, for the second stage, the device transconductance parameters are

$$K_{P2} = k_P^{'} \frac{W_{P2}}{L_{P2}} = 13.4\mu A / V^2 \left(\frac{35}{3} \right) = 0.16 mA / V^2$$

and

$$K_{N2} = k_N^{'} \frac{W_{N2}}{L_{N2}} = 33.4\mu A / V^2 \left(\frac{15}{3} \right) = 0.16 mA / V^2 \cdot$$

The load capacitance seen by the second stage is the input capacitance for the *third* stage,

$$\begin{aligned} C_{L2} &= C_{gP3} + C_{gN3} = C_{ox}\left(W_{P3}L_{P3} + 2W_{P3}L_{OV}\right) + C_{ox}\left(W_{N3}L_{N3} + 2W_{N3}L_{OV}\right) \\ &= 0.575fF\left[(90)(3) + 2(90)(0.2)\right] + 0.575fF\left[(35)(3) + 2(35)(0.2)\right] = 240fF\ . \end{aligned}$$

The propagation delay for the second stage is therefore

$$t_{P2} = \frac{C_{L1}\Gamma}{K_1} = \frac{\left(240 \times 10^{-15}F\right)\left(0.272V^{-1}\right)}{0.16 \times 10^{-3} A / V^2} = 0.41ns\,.$$

For the third stage, the device transconductance parameters are

$$K_{P3} = k_P^{'} \frac{W_{P3}}{L_{P3}} = 13.4\mu A / V^2 \left(\frac{90}{3} \right) = 0.4 mA / V^2$$

and

$$K_{N3} = k_N^{'} \frac{W_{N2}}{L_{N2}} = 33.4\mu A / V^2 \left(\frac{35}{3} \right) = 0.4 mA / V^2 \cdot$$

The load capacitance seen by the third stage is the *external load*,

$$C_{L3} = 15pF \cdot$$

The propagation delay for the third (output) stage is therefore

$$t_{P2} = \frac{C_{L1}\Gamma}{K_1} = \frac{\left(15 \times 10^{-12}F\right)\left(0.272V^{-1}\right)}{0.4 \times 10^{-3} A / V^2} = 10.2ns \cdot$$

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,

$$t_P = t_{P1} + t_{P2} + t_{P3} = 0.67ns + 0.41ns + 10.2ns = 11.3ns\ .$$

Therefore, the output stage accounts for 90% of the overall propagation delay.

6.18 Pseudo NMOS Circuits

Using CMOS fabrication technology, it is possible to implement NMOS-type* circuitry using a passively driven p-channel MOSFET load. This type of logic circuitry is called "pseudo NMOS." There are two notable advantages of pseudo NMOS circuits over conventional NMOS logic, which uses depletion-type as well as enhancement-type n-MOS transistors. First, pseudo NMOS requires significantly fewer transistors than CMOS for the implementation of certain logic functions with potential savings in chip area as well as reductions in the total switched load capacitance. Second, pseudo NMOS only requires the fabrication of enhancement type n-channel MOSFETs, so it is compatible with CMOS processing technology.

A pseudo NMOS inverter is depicted in Figure 6.58. The circuit is identical to an NMOS inverter with the exception of the load, which is a passively driven enhancement type p-channel MOSFET. Because the p-MOS transistor has its gate grounded, it will always conduct. If a low input is applied, the n-MOS device is cutoff whereas the p-MOS device is linear, so the output goes high. If a high input is applied, the n-MOS transistor is linear, whereas the p-MOS transistor is saturated. This causes the output to go low, but in this logic state, there is steady current flow accompanied by significant static dissipation.

Pseudo NMOS allows the realization of arbitrary logic functions by the addition of transistors to the pull-down network, as shown in Figure 6.59. Here, X is a general input vector comprising M scalar inputs. Whereas a CMOS circuit with the same fan-in would require M p-MOS transistors, the pseudo NMOS circuit only uses a single passively driven pull-up device, thus eliminating (M – 1) p-MOS transistors from the layout.

For the pull-down network, the circuit design and scaling principles are identical to those for the pull-down network in CMOS. As an example, consider the three-way NOR (NOR3) gate shown in Figure 6.60. Each electrical path from the output to ground includes just one n-channel MOSFET; hence, these transistors need not be scaled compared with the inverter, nor does the p-channel load transistor require scaling. This contrasts with the CMOS NOR3 circuit, in which there are three series p-channel transistors and each should be scaled up in width by a factor of three. Clearly, the pseudo NMOS circuitry offers potential savings in chip area as well as reductions in the total switched capacitance.

Implementation of the XOR2 function is rather efficient in pseudo NMOS, as shown in Figure 6.61. Generally, the realization of a logic function involving M inputs requires (M + 1) transistors in pseudo NMOS but 2M 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 significant

* NMOS is a logic family based on circuits using only n-MOS devices. Actively driven enhancement-type transistors are used for the pull-down circuitry, whereas a single, passively driven depletion-type transistor is used in the pull-up circuitry.

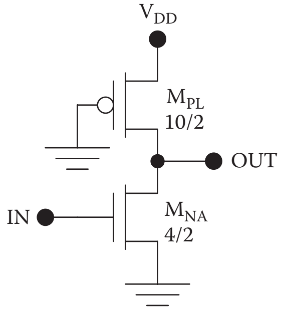


FIGURE 6.58
Pseudo NMOS inverter.

advantage. The disadvantage of pseudo NMOS is the static power dissipation: under output low conditions, a steady DC will flow in the p-channel load and pull-down network.

6.19 Scaling of CMOS

Over the past three decades, scaling of CMOS devices, the systematic reduction of transistor dimensions from one generation to the next, has yielded tremendous gains in chip performance and functionality. Scaling has allowed the industry to keep pace with Moore's law by enabling a reduction in transistor dimensions and therefore area. Device scaling also reduces the parasitic capacitances while increasing the transconductance, thereby improving

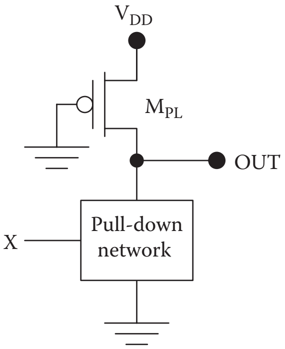


FIGURE 6.59
General pseudo NMOS logic gate.

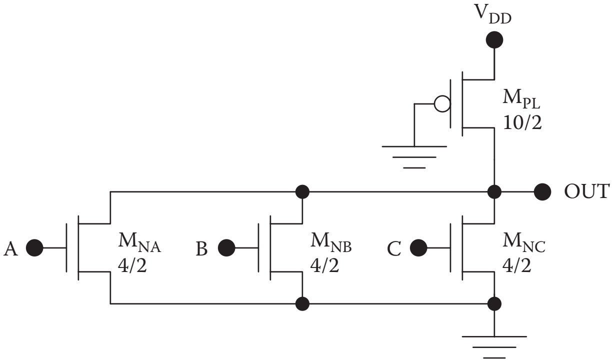


FIGURE 6.60
Pseudo NMOS NOR3 gate.

circuit speed. Many distinct approaches to scaling can be undertaken, but here we will consider two such approaches: *full scaling* and *constant voltage scaling*.

6.19.1 Full Scaling of CMOS

Full scaling involves the scaling of all dimensions and voltages by the same factor $1/\kappa$, where κ is greater than one. For example, if a scaling factor of $1/\sqrt{2}$ is used $\left(\kappa = \sqrt{2}\right)$, then the packing density in transistors per square

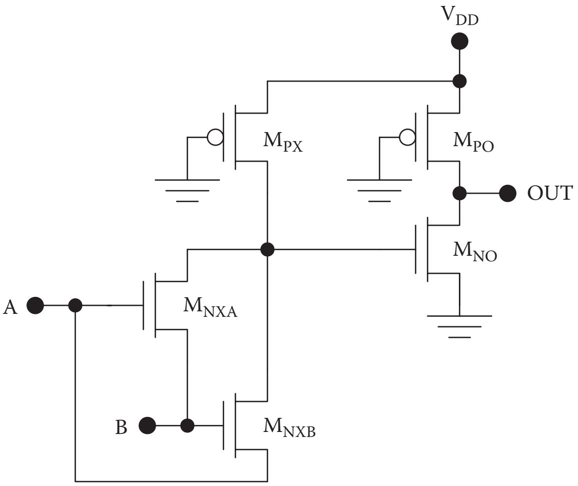


FIGURE 6.61
Pseudo NMOS XOR gate.

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 6.4 provides a summary of CMOS full scaling, including the scaled quantities and the resulting changes in circuit characteristics. Because the gate capacitance of each MOS transistor is reduced by $1/\kappa$, the input capacitance for each CMOS logic circuit scales by $1/\kappa$ as well. The device transconductance parameters increase by κ while the voltages are scaled by $1/\kappa$ so the propagation delay with fixed fan-out scales as $1/\kappa$. Even if the switching frequencies are increased by κ to take advantage of the reduced delays, the switching power dissipation decreases per gate by the reduction in the supply voltage. Therefore, although the packing density (gates per square centimeter) increases as κ^2, the power density (W per centimeter) stays fixed.

6.19.2 Constant Voltage Scaling of CMOS

Another possible approach to scaling of CMOS is called *constant voltage scaling*. This involves the scaling of all dimensions by the factor $1/\kappa$, while all voltages are kept constant. Table 6.5 summarizes the scaled quantities and the resulting changes in circuit characteristics. Constant voltage scaling is convenient in that it does not require a change of supply voltage, and it provides benefits in terms of switching speed performance. The drawback of this approach is a dramatic increase in the power density.

TABLE 6.4

Full Scaling of CMOS

Parameter	**Relationship**	**Scales by**
L, W, t_{OX}, x_j, L_{OV}		$1/\kappa$
V_{DD}, V_{TN}, V_{TP}		$1/\kappa$
Na, Nd		κ
C_{OX}	ε_{ox}/t_{ox}	κ
Cg	$C_{ox}(WL+2WL_{OV})$	$1/\kappa$
K_N, K_P	$\mu_n C_{ox} W_N / L_N$, $\mu_p C_{ox} W_P / L_P$	κ
t_P (fixed C_L)	$\propto C_L/(V_{DD}K)$	1
t_P (fixed fan-out)	$\propto C_g/(V_{DD}K)$	$1/\kappa$
Clock frequency f	$\propto 1/t_P$	κ
P (fixed fan-out)	$\propto fC_g V_{DD}^2$	$1/\kappa^2$
Packing density D		κ^2
Power density	$P \cdot D$	1

TABLE 6.5

Constant Voltage Scaling of CMOS

Parameter	Relationship	Scales by
L, W, t_{OX}, x_j, L_{OV}		$1/\kappa$
V_{DD}, V_{TN}, V_{TP}		1
N_a, N_d		κ^2
C_{OX}	$\varepsilon_{ox} / t_{ox}$	κ
C_g	$C_{ox}\left(WL + 2WL_{OV}\right)$	$1/\kappa$
K_N, K_P	$\mu_n C_{ox} W_N / L_N$, $\mu_p C_{ox} W_P / L_P$	κ
t_P (fixed C_L)	$\propto C_L / \left(V_{DD} K\right)$	$1/\kappa$
t_P (fixed fan-out)	$\propto C_g / \left(V_{DD} K\right)$	$1/\kappa^2$
clock frequency f	$\propto 1/t_P$	κ^2
P (fixed fan-out)	$\propto fC_g V_{DD}^2$	κ
packing density D		κ^2
power density	$P \cdot D$	κ^3

6.20 Latch-Up in CMOS

In integrated form, complementary pairs of n-MOSFETs and p-MOSFETs contain parasitic bipolar junction transistors that combine together in a pnpn structure, or *thyristor*. It is possible for the parasitic thyristor to latch on, effectively shorting V_{DD} to ground. Because of the large resulting current, this *latch-up* condition is generally destructive and must be avoided by careful circuit layout and process design.

Figure 6.62 shows the physical structure of a CMOS inverter with its parasitic bipolar transistors. (An n-well process has been assumed.) The source and drain of the p-MOSFET form two emitters of a pnp bipolar transistor Q_1 with its base in the n-well and its collector in the p-substrate. The source and drain of the n-MOSFET form two emitters of a parasitic npn bipolar transistor Q_2 with its base in the p-substrate and its collector in the n-well. R_{well} and R_{sub} are series resistances in the n-well and substrate, respectively.

Figure 6.63 shows the equivalent circuit diagram for a CMOS inverter including its parasitic bipolar transistors. Under normal static conditions, both bipolar transistors are cutoff because all of the base-emitter junctions experience zero or reverse bias. However, the application of a small and temporary base current to either bipolar transistor can result in a destructive latch-up condition because of the positive feedback connection of these two bipolar transistors. For example, if a small base current is applied to Q_1, the resulting collector current in Q_1 will provide base drive to Q_2. The collector

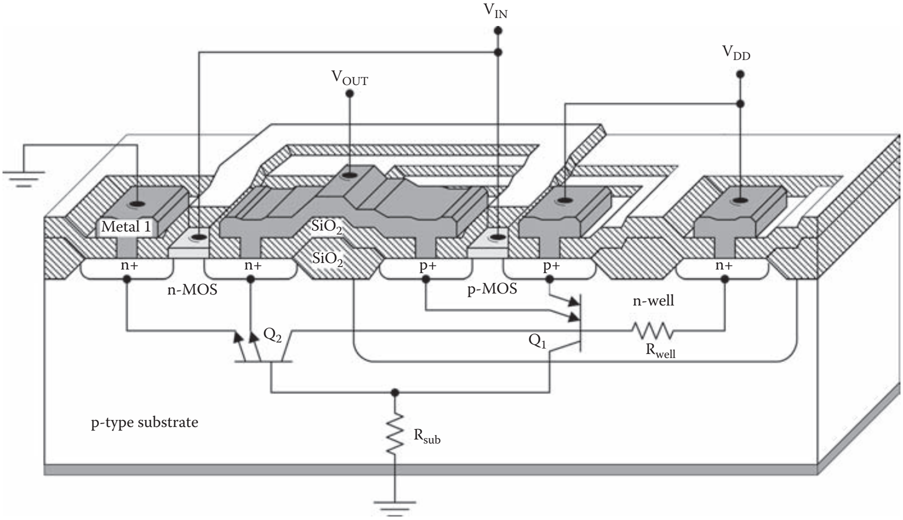


FIGURE 6.62
Physical structure of a CMOS inverter showing the parasitic bipolar transistors.

current in Q_2 will drive the base of Q_1, so that both devices can continue to conduct even after the original current stimulus has been removed. The simultaneous conduction of Q_1 and Q_2, which form a pnpn thyristor, gives rise to a destructive short from V_{DD} to ground. The transient current necessary to trigger the parasitic thyristor may flow in response to a number of conditions, including the power-up transient, noise on the ground and V_{DD} lines, or absorption of alpha particles or cosmic rays.

Latch-up may be prevented by minimizing the resistances R_{well} and R_{sub}, thus limiting the base-to-emitter bias voltages on the parasitic transistors, or by reducing the current gains of Q_1 and Q_2, or by some combination thereof. The current gains of the parasitic transistors can be reduced by increasing the base widths. This in part dictates the minimum spacing between the active region of the n-MOS transistor and the n-well of the p-MOS transistor. However, device scaling of CMOS tends to reduce both base widths and aggravate the problem. Heavier doping can not only reduce the well and substrate resistances but can also reduce the current gains of the parasitic transistors.

6.21 SPICE Demonstrations

For the purpose of illustration, simulations were performed using Cadence Capture CIS 10.1.0 PSpice (Cadence Design Systems). The level 1 MOS transistor

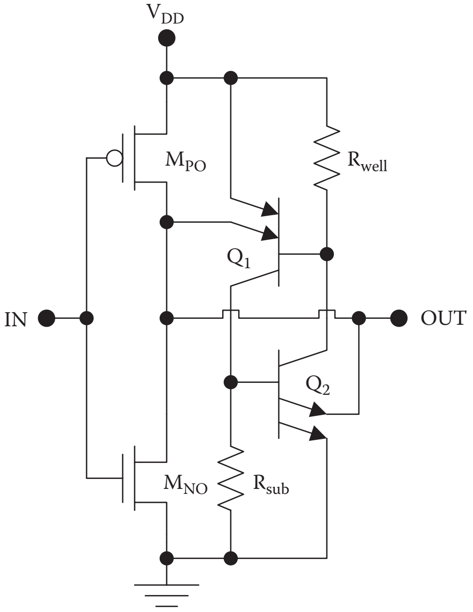


FIGURE 6.63
CMOS inverter equivalent circuit including parasitic bipolar transistors.

model parameters given in Tables 6.6 and 6.7 were used unless otherwise noted. The process transconductance parameters were calculated assuming an oxide thickness of 9 nm. For n-MOSFETS,

$$KP = \frac{(3.9)\left(8.85\times10^{-14}\,F/cm\right)\left(580cm^2V^{-1}s^{-1}\right)}{9\times10^{-7}\,cm} = 222\mu A/V^2, \quad (6.115)$$

TABLE 6.6
n-MOS Level 1 SPICE Parameters

Parameter	Value	Units
KP	222u	A/V^2
VTO	0.5	V
GAMMA	0.15	V$^{1/2}$
PHI	0.7	V
LAMBDA	0.05	
TOX	9n	m
NSUB	1E16	cm^{-3}
UO	580	cm^2/Vs
CGSO	0.38n	F/m
CGDO	0.38n	F/m

TABLE 6.7

p-MOS Level 1 SPICE Parameters

Parameter	Value	Units
KP	88u	A/V^2
VTO	–0.5	V
GAMMA	0.15	V$^{1/2}$
PHI	0.7	V
LAMBDA	0.05	
TOX	9n	m
NSUB	1E16	cm^{-3}
UO	230	cm^2/Vs
CGSO	0.38n	F/m
CGDO	0.38n	F/m

and for p-MOSFETS,

$$KP = \frac{(3.9)\left(8.85\times10^{-14}F/cm\right)\left(230cm^2V^{-1}s^{-1}\right)}{9\times10^{-7}cm} = 88\mu A/V^2 . \tag{6.116}$$

The overlap capacitances per unit gate width were determined with the assumption that $L_{OV} = 0.1\mu m$:

$$CGSO = \frac{(3.9)\left(8.85\times10^{-14}F/cm\right)\left(0.1\times10^{-4}cm\right)}{9\times10^{-7}cm} \tag{6.117}$$

$$= 3.8pF/cm = 0.38nF/m$$

and

$$CGDO = \frac{(3.9)\left(8.85\times10^{-14}F/cm\right)\left(0.1\times10^{-4}cm\right)}{9\times10^{-7}cm} \tag{6.118}$$

$$= 3.8pF/cm = 0.38nF/m.$$

The body effect coefficient was calculated from

$$GAMMA = \frac{\sqrt{2q\varepsilon_{Si}N_a}}{C_{ox}}$$

$$= \frac{\sqrt{2\left(1.602\times10^{-19}C\right)(11.9)\left(8.85\times10^{-14}F/cm\right)\left(10^{16}cm^{-3}\right)}}{(3.9)\left(8.85\times10^{-14}F/cm\right)/9\times10^{-7}cm}$$

$$\approx 0.15V^{1/2}.$$

SPICE Example 6.1 Voltage Transfer Characteristic

The voltage transfer characteristic was determined for the symmetric CMOS inverter shown in Figure 6.64 using a DC sweep of V_{IN} with a step size of 0.01 V. The resulting characteristic appears in Figure 6.65, with critical voltages $V_{IL} = 1.16$ V, $V_M = 1.25$ V, and $V_{IH} = 1.34$ V. The noise margins are both equal to 1.16 V for this symmetric circuit.

SPICE Example 6.2 Voltage Transfer Characteristic with V_{DD} as a Parameter

Voltage transfer characteristics were determined for the symmetric inverter of Figure 6.66 with supply voltages of 1.0, 1.5, 2.0, and 2.5 V, using a parametric

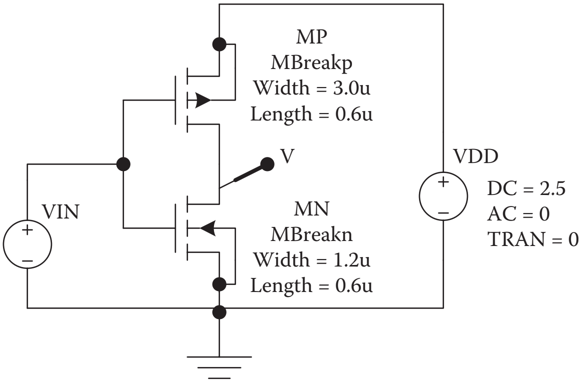


FIGURE 6.64
Symmetric CMOS inverter for the determination of the voltage transfer characteristic.

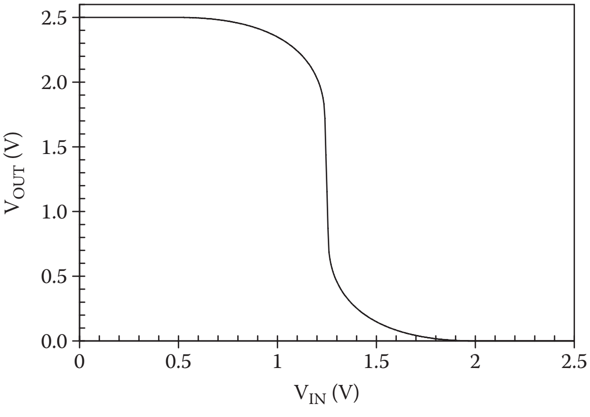


FIGURE 6.65
Voltage transfer characteristic for the symmetric CMOS inverter of Figure 6.64.

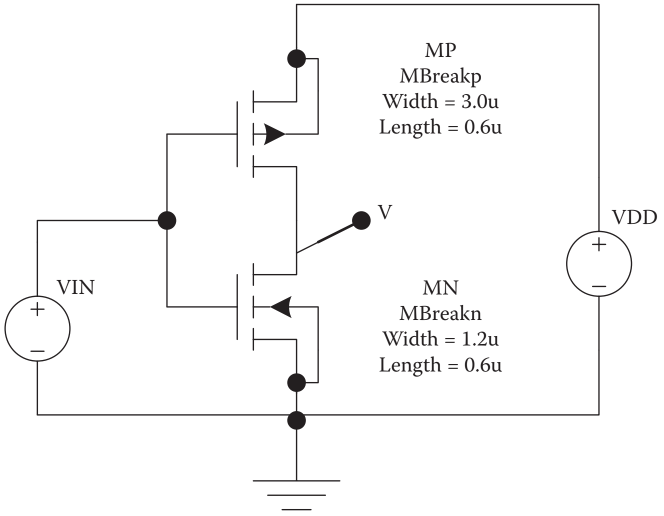


FIGURE 6.66
Symmetric CMOS inverter for the determination of the voltage transfer characteristics with V_{DD} as a parameter.

sweep of V_{DD}. The results of Figure 6.67 show that the transfer characteristic becomes more abrupt as the supply voltage is decreased, when the threshold voltages are held constant. In fact, if $V_{DD} \le V_{TN} + |V_{TP}|$, no crossover current flows, resulting in an extremely sharp output transition (see the case of V_{DD} = 1.0V).

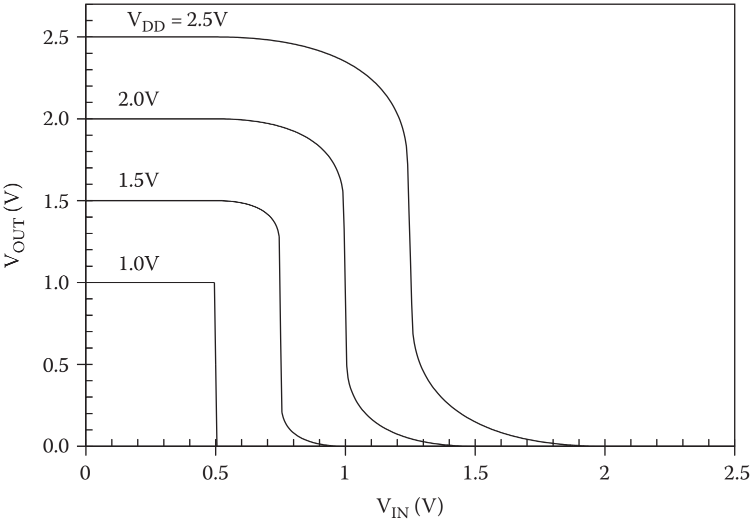


FIGURE 6.67
Voltage transfer characteristic for the symmetric CMOS inverter of Figure 6.66.

SPICE Example 6.3 Voltage Transfer Characteristics for Asymmetric Inverters

Voltage transfer characteristics were determined for the asymmetric inverters as shown in Figure 6.68. The width ratio for the MOS transistors W_P/W_N was set to 0.5, 2, and 8. The results of Figure 6.69 show that increasing the W_P/W_N ratio moves the characteristic to the right; the switching thresholds are 0.96, 1.21, and 1.46 V with $W_P/W_N = 0.5$, 2, and 8, respectively.

SPICE Example 6.4 Crossover Current

The crossover current was determined as a function of the input voltage using a DC sweep and the symmetric CMOS inverter of Figure 6.70. The results in Figure 6.71 indicate a peak crossover current of 132 μA occurring at $V_{DD} \le V_{TN} + |V_{TP}|$.

SPICE Example 6.5 Influence of λ on the Crossover Current

Here, the crossover current was determined as a function of the input voltage for three different values of the channel length modulation parameter: $\lambda = 0$, 0.05, and 0.1. It was assumed that $\lambda_N = \lambda_P = \lambda$, and the circuit of Figure 6.72 was used. The SPICE results shown in Figure 6.73 demonstrate that the channel length modulation parameter has a relatively weak effect on the crossover current characteristic. If an approximate characteristic is calculated using $\lambda = 0$, the maximum error (occurring at the current peak) amounts to about 10% for typical CMOS circuits.

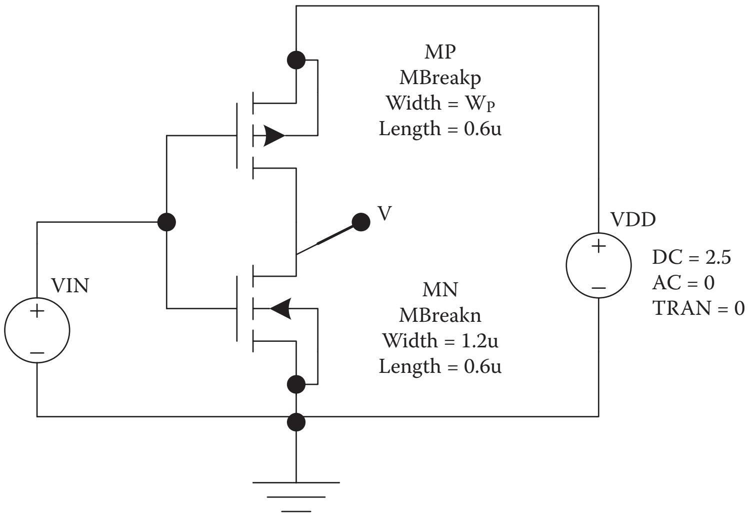


FIGURE 6.68
Asymmetric CMOS inverter for the determination of the voltage transfer characteristics with different values of the transistor width ratio W_P/W_N.

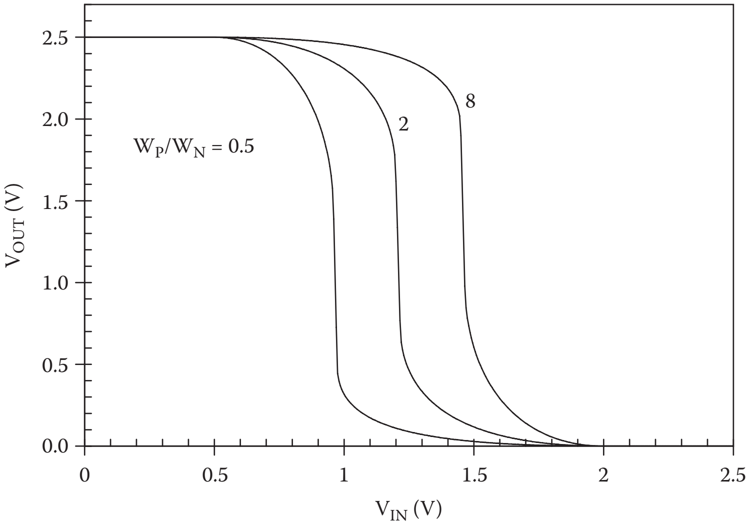


FIGURE 6.69
Voltage transfer characteristics for asymmetric CMOS inverters with W_P/W_N width ratios of 0.5, 2, and 8.

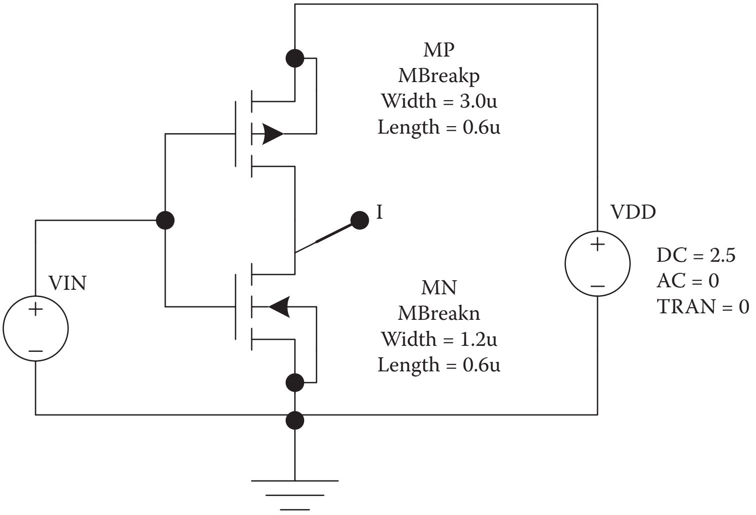


FIGURE 6.70
Symmetric CMOS inverter for the determination of the crossover current.

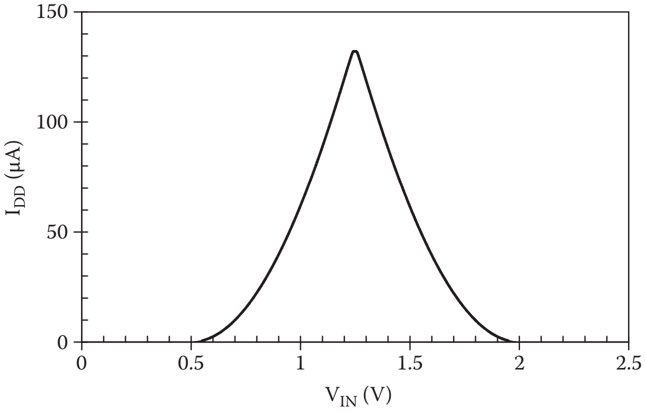


FIGURE 6.71
Crossover current I_{DD} as a function of input voltage for the symmetric CMOS inverter of Figure 6.70.

SPICE Example 6.6 Propagation Delays

A transient simulation was used to determine the propagation delays for the symmetric CMOS inverter of Figure 6.74 with a 1 pF load. The pulse source was set up with V1 = 0, V2 = 2.5 V, TD = 0, TR = 0, TF = 0, PW = 10 ns, and PER = 20 ns. The results in Figure 6.75 show that the symmetric inverter has equal propagation delays $t_{PLH} = t_{PHL} = 1.36ns$.

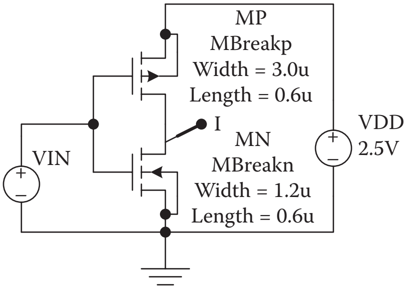


FIGURE 6.72
Symmetric CMOS circuit used for the determination of the crossover current as a function of the input voltage, with λ as a parameter.

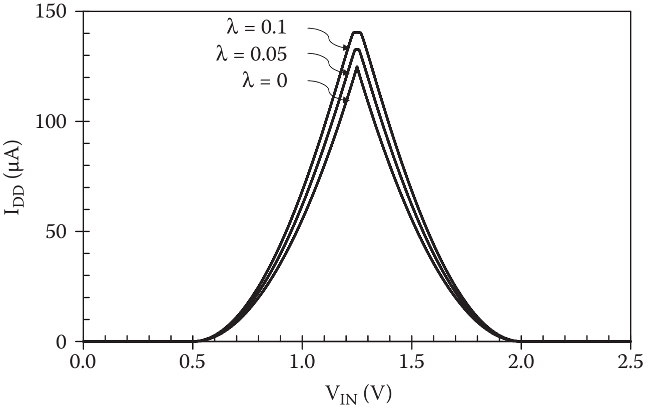


FIGURE 6.73
Crossover current as a function of input voltage with channel length modulation constant as a parameter, for the symmetric CMOS inverter of Figure 6.72.

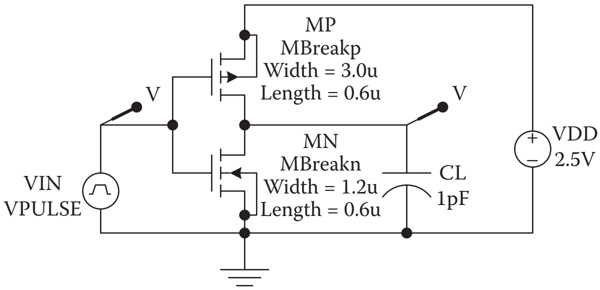


FIGURE 6.74
Symmetric CMOS circuit with a 1 pF load for the determination of the transient response.

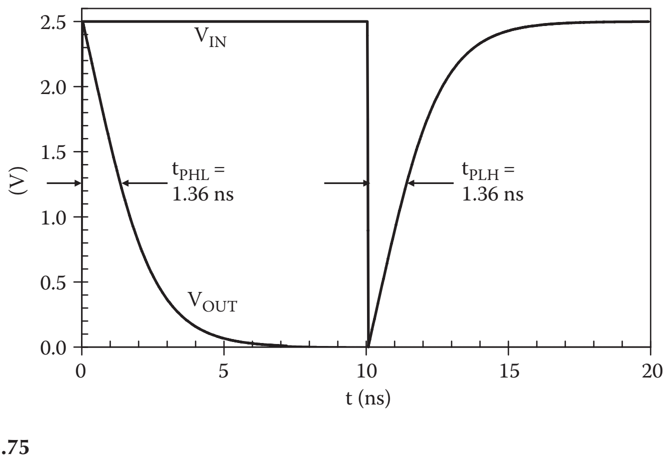


FIGURE 6.75
Transient response for the symmetric CMOS inverter of Figure 6.74.

SPICE Example 6.7 Propagation Delays in an Asymmetric Inverter

For the asymmetric CMOS inverter of Figure 6.76, a series of transient simulations was performed using load capacitances of 50, 100, 200, 300, 400, and 500 fF. In each case, both propagation delays were determined, and the resulting characteristics are plotted in Figure 6.77. It can be seen that the propagation delays are unequal for the asymmetric inverter having equal transistor widths, and $t_{PLH} \approx 2.5 t_{PHL}$ for the case of the long-channel devices considered here. The slopes of the characteristics are $\partial t_{PLH} / \partial C_L = 3300\Omega$ and $\partial t_{PHL} / \partial C_L = 1300\Omega$.

SPICE Example 6.8 CMOS Ring Oscillator

The transient response was determined for a three-stage CMOS ring oscillator with a 500 fF load at each stage as shown in Figure 6.78. The individual gates are the same as the circuit used in SPICE Example 6.7, and the expected propagation

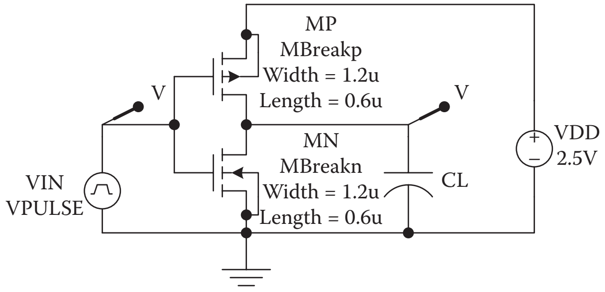


FIGURE 6.76
Asymmetric CMOS inverter for the determination of the propagation delays for various values of load capacitance.

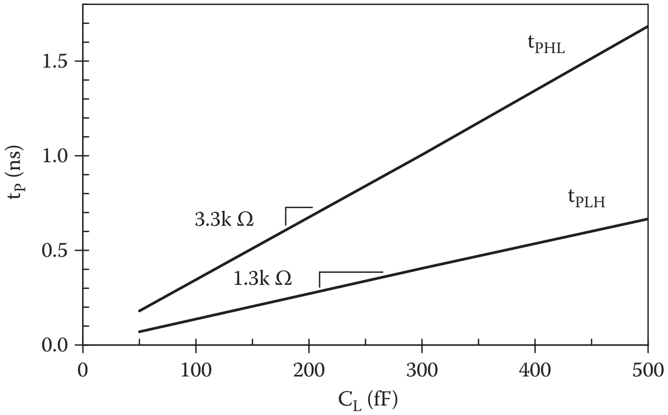


FIGURE 6.77
Propagation delays versus the load capacitance, for the asymmetric CMOS inverter of Figure 6.76.

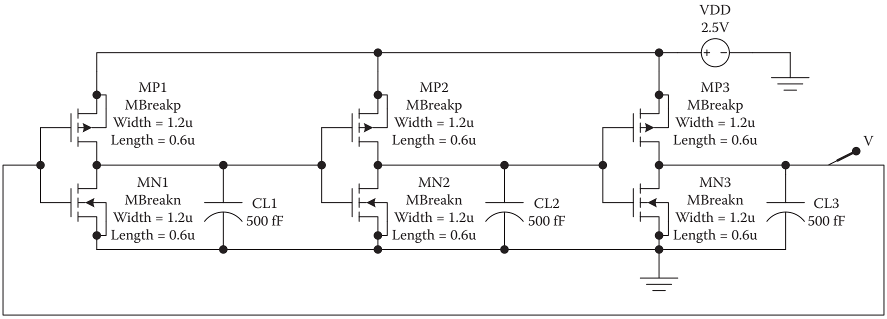


FIGURE 6.78
Three-stage CMOS ring oscillator for the determination of the transient response.

delays are $t_{PLH} \approx 1.68ns$ and $t_{PHL} \approx 0.67ns$. The expected frequency of oscillation is therefore $1/\left[3(1.68ns + 0.67ns)\right] \approx 140MHz$. As seen in the transient simulation results of Figure 6.79, oscillations built up after ~80 ns and exhibited a period of 13 ns, corresponding to a frequency of 77 MHz. This difference is attributable to the fact that the inverters in the ring oscillator all experience long rise and fall times at their inputs, whereas the propagation delays in SPICE Example 6.7 were determined with abrupt input transitions.

6.22 Summary

CMOS digital circuits, constructed using complementary pairs of p-MOS and n-MOS transistors, exhibit near-ideal voltage transfer characteristics, minimal DC dissipation, high packing density, and high speed. The voltage transfer characteristic for a CMOS inverter may be calculated by equating the drain currents in the n-MOS and p-MOS devices; this reveals five regimes based on the modes of operation for the two devices. For a CMOS inverter with abrupt input voltage transitions and a lumped load capacitance, the low-to-high propagation delay is proportional to the load capacitance and inversely proportional to the p-MOS device transconductance parameter. The high-to-low propagation delay is proportional to the load capacitance and inversely proportional to the n-MOS device transconductance parameter. The power dissipation is usually dominated by the capacitance switching power and is proportional to the load capacitance and the square of the supply voltage.

In the case of CMOS implemented with short-channel MOS transistors ($L < 0.5\mu m$ with typical supply voltages), it is necessary to account for carrier velocity saturation. Approximate equations for the propagation delays

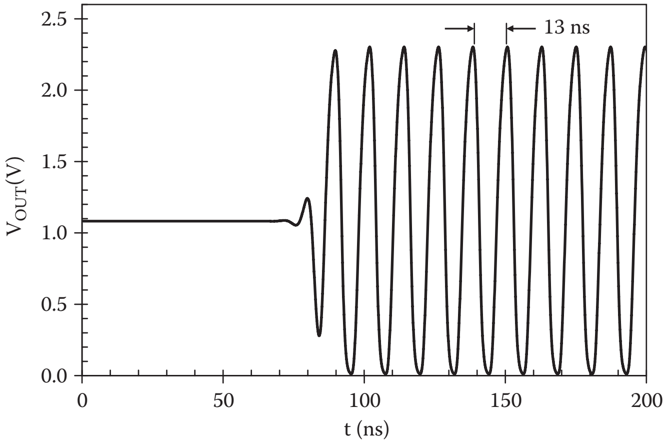


FIGURE 6.79
Transient response for the three-stage CMOS ring oscillator of Figure 6.78.

have been presented for the case of short-channel CMOS. In this case, t_{PLH} and t_{PHL} are proportional to the load capacitance as in long-channel CMOS, but these delays are inversely proportional to the device widths rather than their aspect ratios.

Logic design in CMOS requires the implementation of a pull-down logic network in n-MOS transistors and a pull-up dual logic network in p-MOS transistors. In a NAND gate, the n-MOS transistors are put in series, whereas the p-MOS transistors are placed in parallel. Complex logic functions may be implemented with combinations of parallel- and series-connected transistors.

6.23 Practical Perspective

For practical perspective articles, see the dynamic website at http://www.engr.uconn.edu/ece/books/ayers.

6.24 Exercises

E6.1. For the CMOS inverter of Figure 6.80, determine the mode of operation for each transistor, the supply current, and the output voltage.

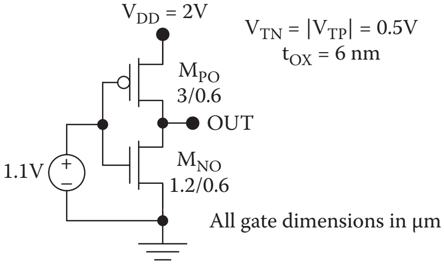


FIGURE 6.80
CMOS inverter for static analysis (see Exercise E6.1).

E6.2. For the CMOS inverter of Figure 6.81, find the mode of operation for each transistor, the supply current, and the output voltage.

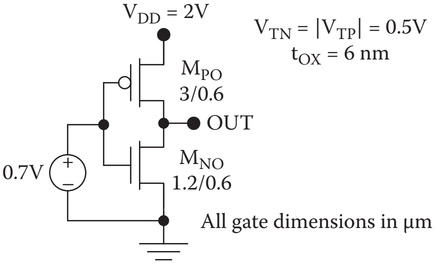


FIGURE 6.81
CMOS inverter for static analysis (see Exercise E6.2).

E6.3. For the asymmetric circuit of Figure 6.82, perform a load curve analysis, that is, produce a graph showing I_{DN} and I_{DP} as functions of V_{OUT} and find the solution at the intersection of the curves.

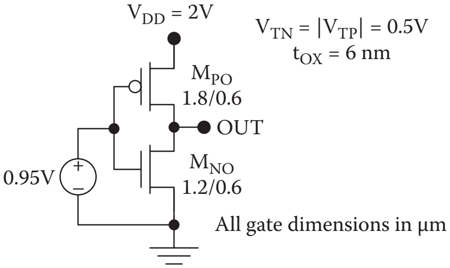


FIGURE 6.82
Inverter circuit for static analysis (see Exercise E6.3).

E6.4. For the circuit of Figure 6.83, perform a load curve analysis with $\lambda_N = \lambda_P = 0.1$, that is, produce a graph showing I_{DN} and I_{DP} as functions of V_{OUT} and indicate the solution on this graph.

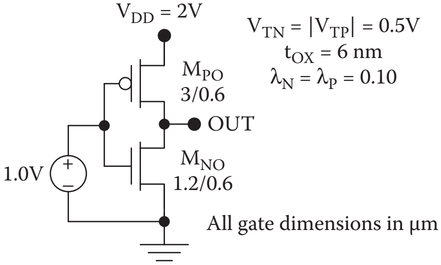

V_{DD} = 2V
M_{PO}
3/0.6
OUT
M_{NO}
1.2/0.6
1.0V
+
–
V_{TN} = |V_{TP}| = 0.5V
t_{OX} = 6 nm
$\lambda_N = \lambda_P$ = 0.10
All gate dimensions in μm

FIGURE 6.83
Inverter circuit for load curve analysis (see Exercise E6.4).

E6.5. Determine and plot the voltage transfer characteristic for the circuit of Figure 6.84. Give the input voltage range for each of the five regimes of this characteristic.

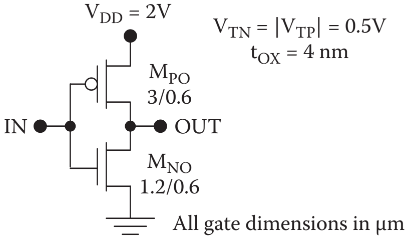


FIGURE 6.84
Inverter for determination of the voltage transfer characteristic (see Exercise E6.5).

E6.6. Determine and plot the voltage transfer characteristics for the circuit of Figure 6.85 with V_{DD} = 2.5, 2.0, and 1.5 V.

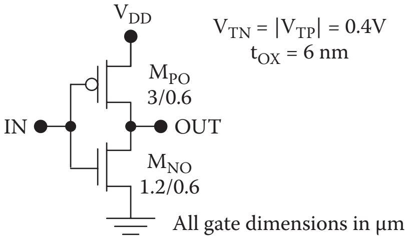


FIGURE 6.85
Inverter for determination of the transfer characteristic (see Exercise E6.6).

E6.7. Find the critical voltages V_{IL}, V_M, and V_{IH} for the inverter of Figure 6.86.

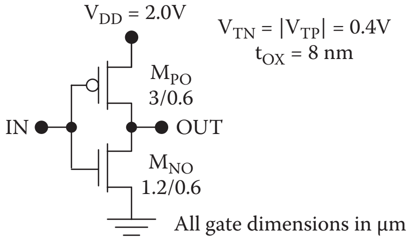


FIGURE 6.86
CMOS inverter for determination of the critical voltages (see Exercise E6.7).

E6.8. Perform a load surface analysis for the circuit in Figure 6.87, that is, produce a surface plot showing I_{DN} and I_{DP} as functions of V_{IN} and V_{OUT}.

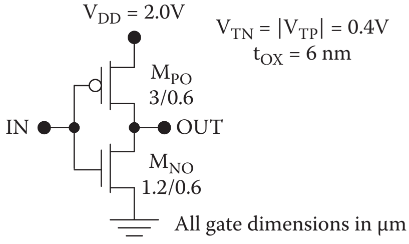


FIGURE 6.87
Inverter circuit for load surface analysis (see Exercise E6.8).

E6.9. Determine and plot the crossover current as a function of V_{IN} for the circuit in Figure 6.88, and find the voltage ranges for each of the four regimes of crossover current.

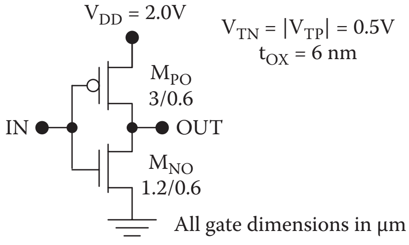


FIGURE 6.88
CMOS circuit for determination of the crossover current (see Exercise E6.9).

E6.10. Plot the crossover current characteristic with V_{DD} as a parameter for $V_{DD} = 2.5$, 2, and 1.5 V, for the circuit of Figure 6.89. How does the peak crossover current depend on the supply voltage?

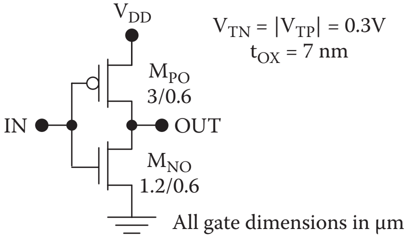


FIGURE 6.89
Inverter for determination of the crossover current with the supply voltage as a parameter (E6.10).

E6.11. For the asymmetric circuit of Figure 6.90, find and plot the crossover current as a function of V_{IN}. Find the range of V_{IN} for each regime of crossover current.

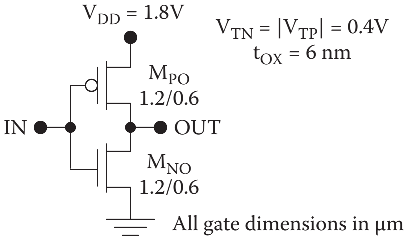


FIGURE 6.90
Asymmetric inverter for determination of the crossover current (see Exercise E6.11).

E6.12. Estimate the propagation delays, the rise time, and the fall time for the circuit in Figure 6.91 assuming the input makes abrupt voltage transitions.

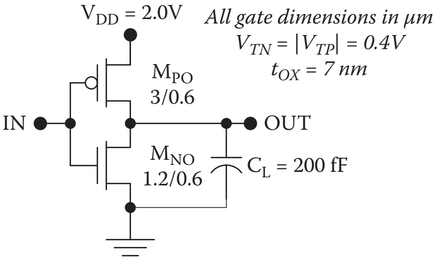


FIGURE 6.91
Loaded inverter for analysis of the delay times (see Exercise E6.12).

E6.13. Estimate the propagation delays for the inverter in Figure 6.92 assuming that the rise time and fall time for the input voltage are both 1 ns.

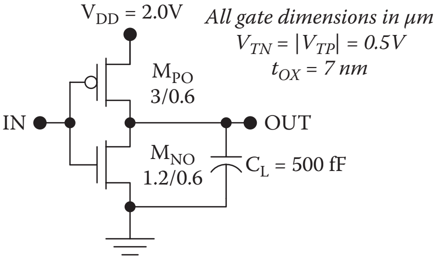


FIGURE 6.92
Loaded inverter for determination of the delay times.

E6.14. Choose the widths of the MOS transistors in Figure 6.93 so that $t_{PLH} \le 100p$ and $t_{PHL} \le 100ps$ with $C_L \le 250f$.

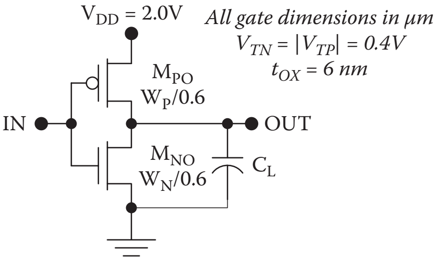


FIGURE 6.93
Inverter circuit with transistor widths to be designed (see Exercise E6.14).

E6.15. Find the input capacitance and estimate the propagation delays for the circuit in Figure 6.94 for the case of three (similar) fan-out gates.

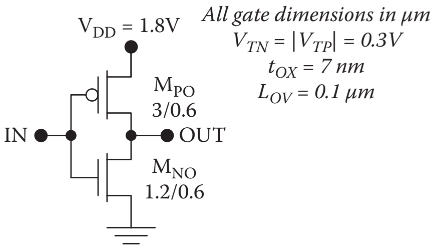


FIGURE 6.94
CMOS inverter for determination of the input capacitance and delay times (see Exercise E6.15).

E6.16. Find the input capacitance and estimate the propagation delays for the asymmetric circuit in Figure 6.95 for the case of three (similar) fan-out gates. Assume that the input voltage makes abrupt transitions.

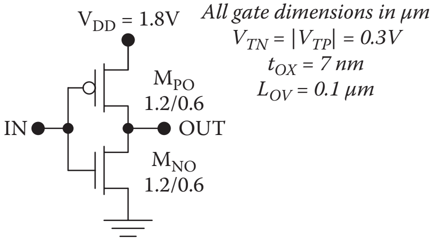


FIGURE 6.95
Asymmetric inverter for determination of the input capacitance and delay times.

E6.17. Using the short-channel MOSFET equations, estimate the propagation delays for the inverter of Figure 6.96 assuming that the input voltage makes abrupt transitions.

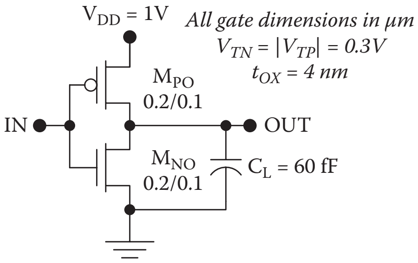


FIGURE 6.96
Short-channel inverter for determination of the delay times (see Exercise E6.17).

E6.18. Find the propagation delays for the circuit of Figure 6.97 assuming that the fan-out is five. Use the short-channel relationships and assume that the input makes abrupt transitions.

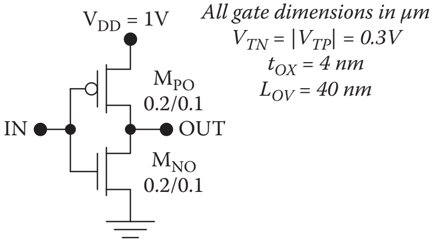


FIGURE 6.97
Short-channel inverter for determination of the delay times (see Exercise E6.18).

E6.19. Find the propagation delays for the circuit of Figure 6.98 assuming that the fan-out is five. Use the short-channel relationships and assume that the input makes abrupt transitions.

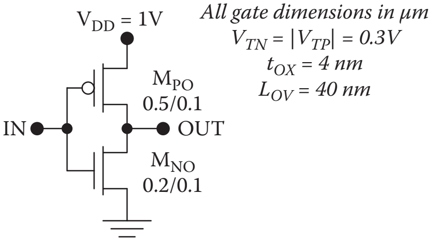


FIGURE 6.98
Short-channel inverter for determination of the delay times (see Exercise E6.19).

E6.20. Three identical inverters are used to make a ring oscillator as shown in Figure 6.99. (1) Find the input capacitance per gate. (2) Estimate the rise and fall times for each of the inverters, assuming a load equal to C_{IN} and abrupt transitions at the input and output. (3) Assuming that each gate experiences input rise and fall times equal to the values determined in the previous part, estimate the propagation delays and the frequency of oscillation for the ring oscillator. (4) Will this frequency be an overestimate or underestimate? Why?

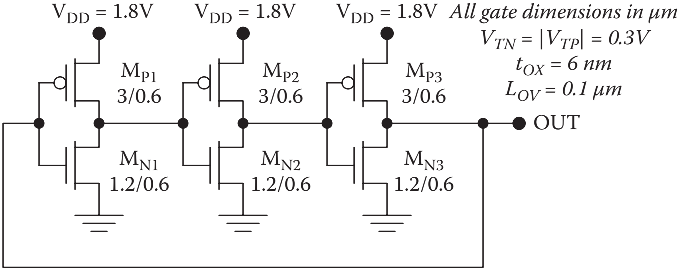


FIGURE 6.99
CMOS ring oscillator (see Exercise E6.20).

E6.21. Find the input capacitance for each gate in the ring oscillator of Figure 6.100. Using an iterative approach, find a consistent set of values for the propagation delays, rise times, and fall times and use this solution to estimate the frequency of oscillation and the capacitance switching dissipation per gate. Use the short-channel MOSFET relationships.

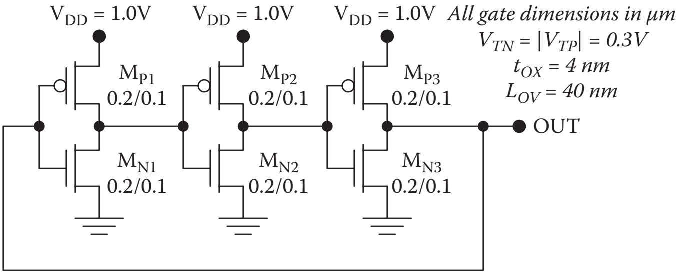


FIGURE 6.100
Three-stage ring oscillator (see Exercise E6.21).

E6.22. A periodic signal is applied to an inverter as shown in Figure 6.101. Estimate the capacitance switching power, the short-circuit power, and the total dynamic dissipation.

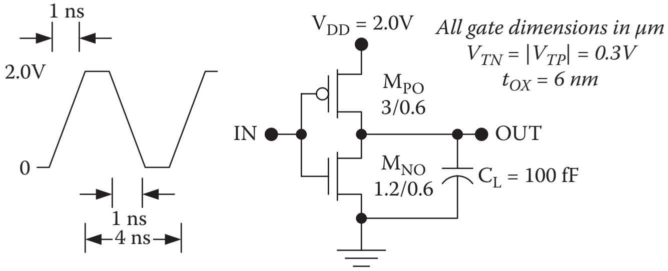


FIGURE 6.101
Loaded inverter for analysis of the dynamic dissipation (see Exercise E6.22).

E6.23. Create the layout design for a CMOS inverter so that $t_{PLH} \leq 100ps$ and $t_{PHL} \leq 100ps$ with $C_L \leq 250fF$. $V_{DD} = 2.0V$, $V_{TN} = |V_{TP}| = 0.3V$, $t_{ox} = 8nm$, and $2X = 0.5\mu m$

E6.24. Create the layout design for a CMOS NAND3 gate using minimum silicon area with $t_{PLH} \leq 100ps$.

E6.25. Create the circuit diagram for a CMOS circuit which implements the function $Y = \overline{(ABC + DE)(F + GH)}$.

For additional exercise problems, see the dynamic website at http://www.engr.uconn.edu/ece/books/ayers.

7

Interconnect

7.1 Introduction

Interconnect refers to the metal wires that make electrical connections between the transistors on the 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 past 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. The 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 the use of longer interconnects on the chip. In such long interconnects, the parasitic resistance must be considered as well as the capacitance. Then 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 above, the parasitics tend to degrade performance. In and of itself, the additional capacitive loading is responsible for increased propagation delays and power. In some digital integrated circuits, the interconnects can account for one-quarter of the total dissipation. On top of this, the parasitic capacitances between interconnects tend to introduce interference, called crosstalk. 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.

In this chapter, we will consider the parasitic resistances, capacitances, and inductances of interconnect. The lumped, distributed, and transmission line models for interconnect will be outlined, and rules will be developed for the determination of which model is most appropriate. Special problems in interconnect design will be discussed from the point of view of minimizing crosstalk and optimizing performance. Then SPICE models for interconnect will be described with some examples.

7.2 Capacitance of Interconnect

The capacitance is the most important parasitic introduced by interconnect wires. Some of this capacitance appears between the interconnect and ground. Another component appears between wires on a single level and is called *interwire capacitance*. There are also capacitances between the interconnect and wires on other levels, the *interlevel capacitance*. Typical VLSI circuits use 8–12 levels of interconnect having complex three-dimensional geometries. Also, process-induced variations in interconnect geometry further complicate modeling. For these reasons, accurate modeling of the capacitive effects is rather complex and computationally intense [5–7]. Therefore, simple models have been developed for the estimation of interconnect capacitances [8].

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 7.1. There are two components of the capacitance between this *microstripline* and the substrate (ground plane). The first is the parallel plate capacitance associated with the parallel field lines directly underneath the interconnect. The second is the fringing field capacitance. The relative contributions of these two components depend on the aspect ratio h/w for the wire. The capacitance per unit length is given by the following [8]:

$$c = \varepsilon_{DI}\left[\frac{\left(w-\dfrac{h}{2}\right)}{t}+\frac{2\pi}{\ln\left(1+\dfrac{2t}{h}+\sqrt{\dfrac{2t}{h}\left(\dfrac{2t}{h}+2\right)}\right)}\right], \text{ for } h/w \le 2, \tag{7.1}$$

and

* The rectangular cross sections for wires are dictated by the deposition, lithography, and etching processes.

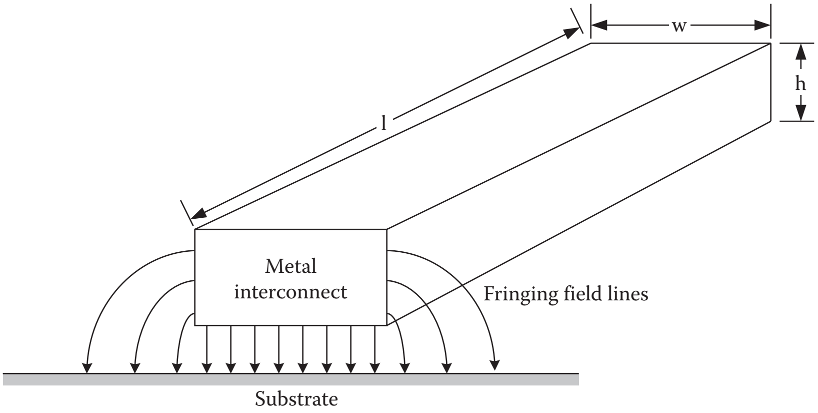


FIGURE 7.1
Estimation of the interconnect capacitance.

$$c = \varepsilon_{DI}\left[\frac{w}{t} + \frac{\pi\left(1 - 0.0543\dfrac{h}{2t}\right)}{\ln\left(1 + \dfrac{2t}{h} + \sqrt{\dfrac{2t}{h}\left(\dfrac{2t}{h} + 2\right)}\right)} + 1.47\right], \text{ for } h/w \geq 2, \qquad (7.2)$$

where ε_{DI} is permittivity of the dielectric, t is thickness of the dielectric between the interconnect and the substrate, w is width of the interconnect, h is the height of the interconnect, and l is the length of the interconnect. The capacitance per unit length can depart significantly from the value predicted for a parallel plate capacitor, $c = \varepsilon_{DI}w/t$, especially for high aspect-ratio wires common in VLSI circuits.

Typical relative permittivities for dielectric materials are shown in Table 7.1. Chemical vapor deposited (CVD) SiO_2 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-κ[*] dielectrics,* such as CVD carbon-doped oxide or spin-on polymers, must be used [9–15].

Example 7.1 Capacitance of Polysilicon Interconnect

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 SiO_2 ($\varepsilon_r = 3.9$).

[*] Sometimes, the relative permittivity is denoted by the Greek letter κ. Therefore, materials with low permittivity are often called "low-κ dielectrics."

TABLE 7.1

Relative Permittivities 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

Solution: Here h/w = 1 so we should use Equation 7.1. The capacitance to the substrate per unit length is

$$c = \varepsilon_{DI}\left[\frac{\left(w-\frac{h}{2}\right)}{t}+\frac{2\pi}{\ln\left(1+\frac{2t}{h}+\sqrt{\frac{2t}{h}\left(\frac{2t}{h}+2\right)}\right)}\right]$$

$$=(3.9)\left(8.85\times10^{-14}F/cm\right)\left[\frac{0.125\times10^{-4}cm}{0.5\times10^{-4}cm}+\frac{2\pi}{\ln\left(1+4+\sqrt{4(4+2)}\right)}\right].$$

$$=1.03\times10^{-12}F/cm=1.03pF/cm$$

This value is representative of a polysilicon wire running over the field oxide in 0.25 µm CMOS. It is much greater than the parallel plate capacitance for the same geometry, which is 0.17 pF/cm. Therefore, the fringing field capacitance is dominant, and the capacitance per unit length is a weak function of the polysilicon width.

Example 7.2 Capacitance of Aluminum Interconnect

Estimate the capacitance to ground per unit length for an aluminum interconnect, 0.25 µm wide and 0.75 µm thick, on a 0.5-µm-thick layer of SiO_2 ($\varepsilon_r = 3.9$).

Solution: Here h/w = 3, so we should use Equation 7.2. The capacitance to ground per unit length is

$$c = \varepsilon_{DI}\left[\frac{w}{t}+\frac{\pi\left(1-0.0543\frac{h}{2t}\right)}{\ln\left(1+\frac{2t}{h}+\sqrt{\frac{2t}{h}\left(\frac{2t}{h}+2\right)}\right)}+1.47\right]$$

$$=(3.9)\left(8.85\times10^{-14}F/cm\right)\left[\frac{0.25\times10^{-4}cm}{0.5\times10^{-4}cm}+\frac{\pi\left(1-0.0543(0.75)\right)}{\ln\left(1+1.33+\sqrt{1.33(1.33+2)}\right)}+1.47\cdot\right.$$

$$=1.38\times10^{-12}F/cm=1.38pF/cm$$

These numbers are representative of a level 1 metal interconnect used in a 0.25 μm CMOS process. As in the previous example, the fringing field capacitance is dominant because of the high aspect ratio for the interconnect. It is possible to neglect the fringing field capacitance only for very wide interconnects with w >> 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.

7.3 Resistance of Interconnect

Consider a metal interconnect with a rectangular cross section as shown in Figure 7.2.

The low-frequency resistance of this interconnect may be calculated as

$$R = \frac{l}{w}\frac{\rho}{h}, \tag{7.3}$$

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

$$r = \frac{\rho}{wh}. \tag{7.4}$$

Typical resistivities for interconnect materials are shown in Table 7.2.

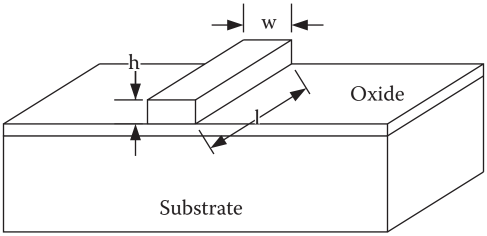


FIGURE 7.2
Interconnect structure for the estimation of the resistance.

TABLE 7.2

Resistivities of 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}

Example 7.3 Resistance of Aluminum Interconnect

Estimate the low-frequency resistance per unit length for an aluminum interconnect, 0.5 μm thick and 0.5 μm wide.

Solution: Assuming the interconnect takes on the bulk resistivity of aluminum,

$$r = \frac{\rho}{wh} = \frac{\left(2.8 \times 10^{-6} \Omega cm\right)}{\left(0.5 \times 10^{-4} cm\right)\left(0.5 \times 10^{-4} cm\right)} = 1.12 k\Omega / cm .$$

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 in Section 7.5.

At high frequencies, the resistance becomes frequency dependent attributable to the *skin effect* [16]. 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 $1/e \approx 0.37$ times the surface value. This skin depth is given by

$$\delta = \sqrt{\frac{\rho}{\pi f \mu}} , \tag{7.5}$$

where f is the frequency, and μ is the permeability of the dielectric.

This phenomenon increases the effective resistance of the interconnect and can be accounted for approximately by assuming that the cross-sectional area is reduced to that within the skin depth as shown in Figure 7.3. Based on this figure, the effective resistance of the interconnect is approximately

$$R = \frac{l\rho}{\delta\left[2w + 2(h - 2\delta)\right]} , \tag{7.6}$$

and the resistance per unit length is

$$r = \frac{\rho}{\delta\left[2w + 2(h - 2\delta)\right]} , \tag{7.7}$$

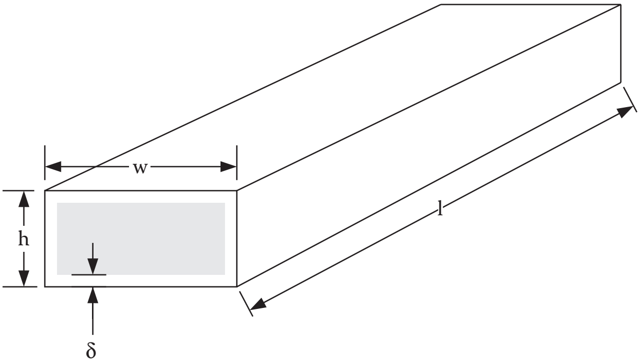


FIGURE 7.3
Skin effect in interconnect.

Example 7.4 Skin Effect in Copper Interconnect

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 both have permeabilities equal to that of a vacuum. At 100 GHz, the skin depth is

$$\delta(Cu)=\sqrt{\frac{\rho}{\pi f\mu}}=\sqrt{\frac{\left(1.7\times10^{-6}\Omega cm\right)}{\pi\left(10^{11}Hz\right)\left(4\pi\times10^{-9}H/cm\right)}}=0.21\times10^{-4}cm=0.21\mu m\,.$$

Therefore, the skin effect would be expected to be important at 100 GHz in copper interconnects having both dimensions greater than twice this value. In practice, the skin effect has not been very important up to the present time because higher-frequency circuits use interconnects with smaller cross sections.

Example 7.5 Resistances of Aluminum and Copper Interconnect

Consider interconnects 1 μm thick and 1 μm wide. Compare the resistances per unit length for aluminum and copper interconnect assuming a frequency of 10 GHz.

Solution: It is assumed that these metals and the dielectric in which they are embedded all have permeabilities equal to that of a vacuum. At 10 GHz, the skin depths are

$$\delta(Cu)=\sqrt{\frac{\rho}{\pi f\mu}}=\sqrt{\frac{\left(1.7\times10^{-6}\Omega cm\right)}{\pi\left(10^{10}Hz\right)\left(4\pi\times10^{-9}H/cm\right)}}=0.66\times10^{-4}cm=0.66\mu m$$

and

$$\delta(Al) = \sqrt{\frac{\rho}{\pi f \mu}} = \sqrt{\frac{\left(2.8\times10^{-6}\Omega cm\right)}{\pi\left(10^{10}Hz\right)\left(4\pi\times10^{-9}H/cm\right)}} = 0.85\times10^{-4}cm = 0.85\mu m\,.$$

Therefore, the skin effect is unimportant here. On a per unit length basis, the approximate interconnect resistances are

$$r(Cu) = \frac{\left(1.7\times10^{-6}\Omega cm\right)}{\left(1.0\times10^{-4}cm\right)\left(1.0\times10^{-4}cm\right)} = 170\Omega/cm$$

and

$$r(Al) = \frac{\left(2.8\times10^{-6}\Omega cm\right)}{\left(1.0\times10^{-4}cm\right)\left(1.0\times10^{-4}cm\right)} = 280\Omega/cm\,.$$

The 37% lower bulk resistivity of copper translates directly to lower parasitic resistances, as long as the bulk resistivity can be achieved in copper thin films.

The preceding analysis is very approximate, because it assumes a sinusoidal current waveform and the abrupt confinement of the current within the skin depth. However, we can conclude that the skin effect can usually be neglected when modeling interconnect for VLSI circuits.

7.4 Inductance of Interconnect

The inductance per unit length for an interconnect may be estimated most easily if the capacitance per unit length is known, using the approximate expression

$$l \approx \frac{\varepsilon_{DI}\mu_{DI}}{c}, \tag{7.8}$$

where ε_{DI} and μ_{DI} are the permittivity and permeability of the dielectric, respectively.

Example 7.7 Inductance of Copper Interconnect

Estimate the inductance per unit length for a copper interconnect, 0.75 μm wide, 0.25 μm thick on a 0.5-μm-thick layer of SiO_2 ($\varepsilon_r = 3.9$).

Solution: From Example 7.2, the capacitance to ground per unit length of this interconnect is 1.38 pF/cm. Therefore, the inductance per unit length is approximately

$$l = \frac{\varepsilon_{DI}\mu_{DI}}{c} = \frac{(3.9)\left(8.85\times10^{-14} F/cm\right)\left(4\pi\times10^{-9} H/cm\right)}{1.38\times10^{-12} F/cm} = 3.2\times10^{-9} H/cm\cdot$$

This inductance is very small and can be neglected under most circumstances. So, whereas the 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 [17–21].

7.5 Modeling Interconnect Delays

In many cases, adequate performance predictions can be made using a lumped capacitance model for the interconnects in a circuit. In some cases, such as polysilicon interconnect, distributed models must be used. Transmission line models may be necessary for extremely long runs of interconnect.

7.5.1 Lumped Capacitance Model

It is reasonable to model short runs of interconnect using a *lumped capacitance model* as shown in Figure 7.4. This approach is valid even for a branching

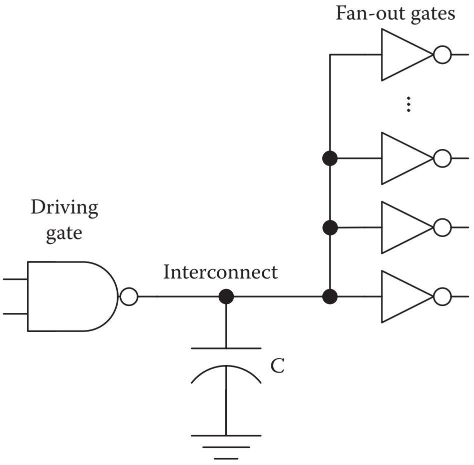


FIGURE 7.4
Use of a lumped capacitance model for interconnect.

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. This in turn increases the propagation delay and the dissipation for the driving gate.

7.5.2 Distributed Models

In longer runs of interconnect, the parasitic resistances becomes important. Therefore, distributed RC models should be used in these situations. Consider a general branching interconnect modeled by an Nth-order distributed network as shown in Figure 7.5.

The Nth-order network exhibits N time constants, and the exact analysis is rather complex. However, it is usually sufficient to consider only the first-order time constant (the *Elmore delay*), which greatly simplifies the analysis [21–24]. Based on this approach, the propagation delay from the driving gate (at node zero) to the ith node is approximately

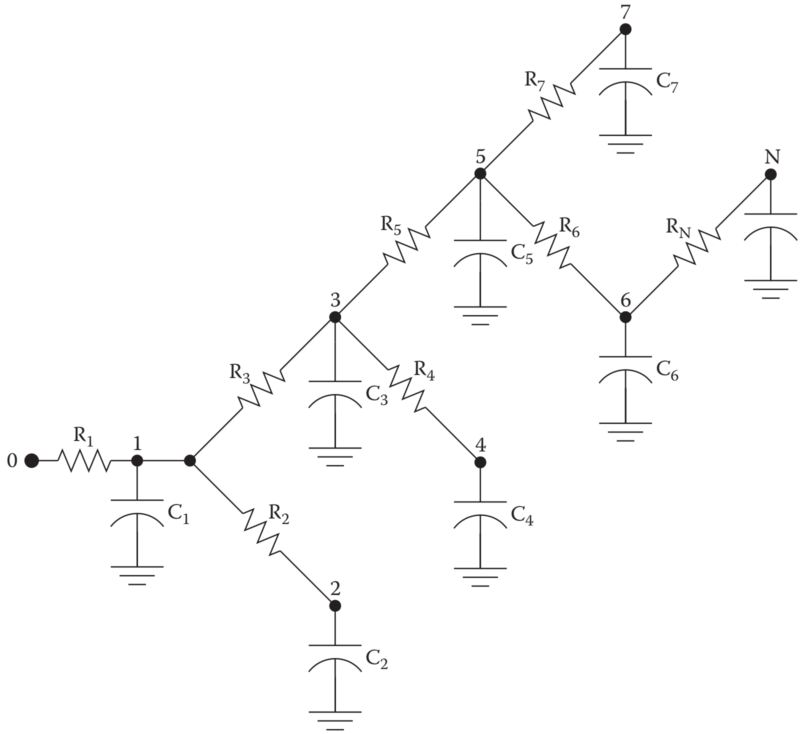


FIGURE 7.5
Distributed RC model for branching interconnect.

$$t_{Pi} \approx \ln(2)\sum_{k=1}^{N} C_k R_{ik} \ , \tag{7.9}$$

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

$$R_{ik} = \sum R_j \in \left[path(0 \to i) \cap \left(path(0 \to k) \right) \right]. \tag{7.10}$$

A special case is the straight (nonbranching) interconnect shown in Figure 7.6.

Here, the Elmore approach yields an approximate end-to-end propagation delay of

$$\begin{aligned} t_P &\approx \ln(2)\left[C_1 R_1 + C_2 \left(R_1 + R_2 \right) + C_3 \left(R_1 + R_2 + R_3 \right) + \ldots \right] \\ &= \ln(2)\sum_{i=1}^{N} C_i \sum_{j=1}^{i} R_j . \end{aligned} \tag{7.11}$$

A situation of special interest is the nonbranching interconnect with evenly distributed resistance and capacitance, as shown in Figure 7.7.

The propagation delay for the evenly distributed RC network, divided into N segments, is

$$t_P \approx \ln(2)\left[\frac{RC}{N^2} + \frac{2RC}{N^2} + \frac{3RC}{N^2} + \ldots + \frac{NRC}{N^2} \right] = \ln(2) RC \frac{N+1}{2N} . \tag{7.12}$$

If the number of segments is increased arbitrarily, the propagation delay asymptotically reaches the limiting value

$$t_P \approx \ln(2) \underset{N \to \infty}{Lim} \left[RC \frac{N+1}{2N} \right] = \ln(2) \frac{RC}{2} = \ln(2) \frac{rcl^2}{2} \ , \tag{7.13}$$

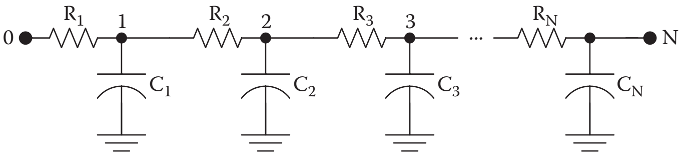


FIGURE 7.6
Distributed RC model for straight (nonbranching) interconnect.

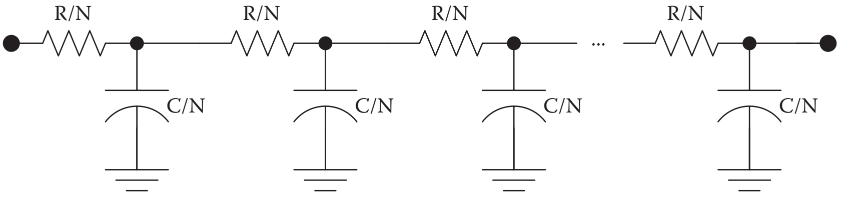


FIGURE 7.7
Straight interconnect with evenly distributed parasitics.

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* [25].

The choice of a lumped capacitance model or the distributed RC model for the interconnect may be made after comparison of 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 follows:

$$l < \sqrt{\frac{t_{P,driver}}{rc\ln(2)/2}} \quad \Rightarrow \quad \text{lumped C model}$$

$$l > \sqrt{\frac{t_{P,driver}}{rc\ln(2)/2}} \quad \Rightarrow \quad \text{distributed RC model}$$

7.5.3 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 it 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

$$v = \frac{c_0}{\sqrt{\varepsilon_r \mu_r}}, \tag{7.14}$$

where c_0 is speed of light in a vacuum (3.00 × 10^{10} cm/s), ε_r is the relative permittivity of the transmission line medium, and μ_r is the relative permeability of the transmission line medium.

Example 7.8 Time of Flight for Interconnect with SiO_2 as the Dielectric

Estimate the time of flight for a signal crossing a 1 cm integrated circuit by a copper transmission line embedded in SiO_2.

Solution: For SiO_2, the relative permittivity is 3.9 and the relative permeability is ~1.0. Thus, the speed of propagation is

$$v = \frac{c_0}{\sqrt{\varepsilon_r \mu_r}} = \frac{3.0 \times 10^{10} cm / s}{\sqrt{(3.9)(1.0)}} = 1.52 \times 10^{10} cm / s \; .$$

The time of flight is

$$t_{flight} = \frac{1cm}{1.52 \times 10^{10} cm / s} = 6.6 \times 10^{-11} s = 66 ps \cdot$$

This time of flight will be comparable with 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.

Another important issue relating to the transmission line model is the characteristic impedance and impedance matching. Consider a general two-wire transmission line as shown in Figure 7.8. The equations governing the general two-wire transmission line are

$$\frac{\partial^2 V}{\partial x^2} = -\left(r + j2\pi fl\right)\left(g + j2\pi fc\right)V \; , \tag{7.15}$$

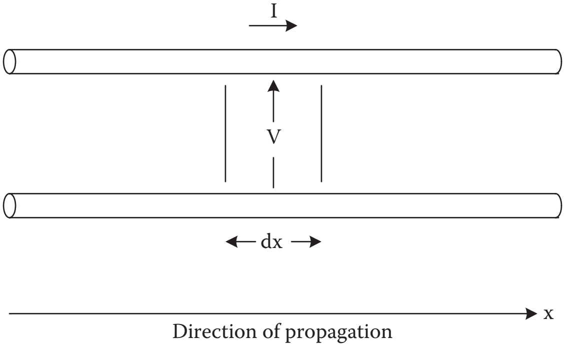


FIGURE 7.8
Two-wire transmission line.

and

$$\frac{\partial^2 I}{\partial x^2} = -\left(r + j2\pi fl\right)\left(g + j2\pi fc\right)I \,, \tag{7.16}$$

where I is the phasor current at the point x, V is the phasor voltage at the point x, r is series resistance per unit length, l is series inductance per unit length, g is parallel conductance per unit length, c is parallel capacitance per unit length, and f is 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

$$\frac{\partial^2 V}{\partial x^2} = lc\frac{\partial^2 V}{\partial t^2} = \frac{1}{v^2}\frac{\partial^2 V}{\partial t^2}\,. \tag{7.17}$$

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,

$$\frac{V}{I} = \sqrt{\frac{l}{c}} = Z_0\,. \tag{7.18}$$

Thus, the remainder of the line appears to have a real impedance Z_0, called the *characteristic impedance* of the transmission line.

Example 7.9 Characteristic Impedance for Aluminum Interconnect

Estimate the characteristic impedance for aluminum interconnect, 0.25 μm wide and 0.75 μm thick, on a 0.5-μm-thick layer of SiO_2.

Solution: The capacitance per unit length is

$$c = \varepsilon_{DI}\left[\frac{w}{t} + \frac{\pi\left(1 - 0.0543\dfrac{h}{2t}\right)}{\ln\left(1 + \dfrac{2t}{h} + \sqrt{\dfrac{2t}{h}\left(\dfrac{2t}{h} + 2\right)}\right)} + 1.47\right]$$

$$= (3.9)\left(8.85\times10^{-14}F/cm\right)\left[\frac{0.25\times10^{-4}cm}{0.5\times10^{-4}cm} + \frac{\pi\left(1 - 0.0543(0.75)\right)}{\ln\left(1 + 1.33 + \sqrt{1.33(1.33+2)}\right)} + 1.47\right]$$

$$= 1.38\times10^{-12}F/cm = 1.38pF/cm$$

The inductance per unit length is approximately

$$l = \frac{\varepsilon_{DI}\mu_{DI}}{c} = \frac{(3.9)\left(8.85\times10^{-14}F/cm\right)\left(4\pi\times10^{-9}H/cm\right)}{1.38\times10^{-12}F/cm} = 3.2\times10^{-9}H/cm\,\cdot$$

The characteristic impedance of the line is therefore

$$Z_0 = \sqrt{\frac{l}{c}} = \sqrt{\frac{3.2\times10^{-9}H/cm}{1.38\times10^{-12}F/cm}} = 48\Omega \cdot$$

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

$$\rho = \frac{R_L - Z_0}{R_L + Z_0} \cdot \tag{7.19}$$

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 driving gate circuit. In terms of the interconnect length l, this rule is as follows:

$$l < \frac{t_{P,driver}c_0}{\sqrt{\varepsilon_r\mu_r}} \quad \Rightarrow \quad \text{distributed RC model}$$

$$l > \frac{t_{P,driver}c_0}{\sqrt{\varepsilon_r\mu_r}} \quad \Rightarrow \quad \text{transmission line model}$$

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, and whereas the simply theory outlined above applies to lossless transmission lines, real interconnects are lossy, causing an attenuation of the signals propagating on them.

Example 7.10 Choice of Interconnect Model

For copper interconnect, 0.5 μm wide and 0.5 μm thick, on a 1.0-μm-thick layer of SiO_2, 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

$$r = \frac{\rho}{wh} = \frac{\left(1.7\times10^{-6}\Omega cm\right)}{\left(0.5\times10^{-4}cm\right)\left(0.5\times10^{-4}cm\right)} = 680\Omega/cm \cdot$$

The capacitance per unit length is

$$c = \varepsilon_{DI}\left[\frac{\left(w-\frac{h}{2}\right)}{t}+\frac{2\pi}{\ln\left(1+\frac{2t}{h}+\sqrt{\frac{2t}{h}\left(\frac{2t}{h}+2\right)}\right)}\right]$$

$$= (3.9)\left(8.85\times10^{-14}F/cm\right)\left[\frac{0.25\times10^{-4}cm}{1.0\times10^{-4}cm}+\frac{2\pi}{\ln\left(1+4+\sqrt{4(4+2)}\right)}\right].$$

$$= 1.03\times10^{-12}F/cm = 1.03pF/cm$$

The lumped capacitance model applies if

$$l<\sqrt{\frac{t_{P,driver}}{rc\ln(2)/2}}=\sqrt{\frac{50\times10^{-12}s}{(680\Omega/cm)\left(1.03\times10^{-12}F/cm\right)\ln(2)/2}}=4500\mu m.$$

The transmission line model applies if

$$l>\frac{t_{P,driver}c_0}{\sqrt{\varepsilon_r\mu_r}}=\frac{\left(50\times10^{-12}s\right)\left(3.0\times10^{10}cm/s\right)}{\sqrt{(3.9)(1.0)}}=7600\mu m.$$

Clearly, the lumped capacitance model will be appropriate in most cases, and it will rarely be necessary to invoke a transmission line model. This could change, however, with the trend toward finer lines and larger die sizes.

7.6 Crosstalk

In a multilevel metallization scheme, the interconnects exhibit parasitic capacitances to ground, 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 *crosstalk* may compromise the integrity of signals and increase the propagation delays for data and address lines.

The worst manifestation of crosstalk occurs in the case of data lines that are run side by side for a long distance, as in a data bus, as shown in Figure 7.9. 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 (which was already discussed in the context of MOSFET capacitances in Chapter 4). Suppose that the victim interconnect makes a low-to-high transition, with

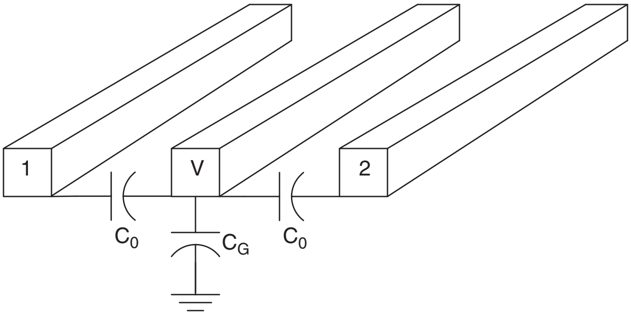


FIGURE 7.9
Parallel interconnects for the consideration of crosstalk.

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

$$C_{eff} = \frac{\Delta Q}{\Delta V} = \frac{V_{DD}C_G + 2V_{DD}C_0}{V_{DD}} = C_G + 2C_0\,, \tag{7.20}$$

where C_G is capacitance to ground for the victim line and C_0 is the capacitance between each pair of neighboring lines.

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

$$C_{eff} = C_G\,. \tag{7.21}$$

Finally, suppose that 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

$$C_{eff} = C_G + 4C_0\,. \tag{7.22}$$

The various permutations are elaborated in Table 7.3. 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, C_G is the actual capacitance to ground for the victim line, and the interwire capacitances to the neighboring wires are each assumed to be C_0.

As a result of crosstalk, 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. This in turn makes performance predictions difficult.

TABLE 7.3

Cross-Talk-Induced Miller Effect in Interconnect

Victim	Neighbor 1	Neighbor 2	C_{eff}
↑	↑	↑	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

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 7.10. This causes the interlevel capacitances to split into many small components, none of which is significant to cause cross-talk problems. Another is to avoid long stretches of parallel lines by creative routing. This approach is facilitated by computer routing tools. Another approach is the insertion of ground and V_{DD} lines between signal lines as illustrated in Figure 7.11. There is no Miller effect with respect to the power rails. Therefore, although the total capacitance to ground is increased somewhat, at least it is predictable.

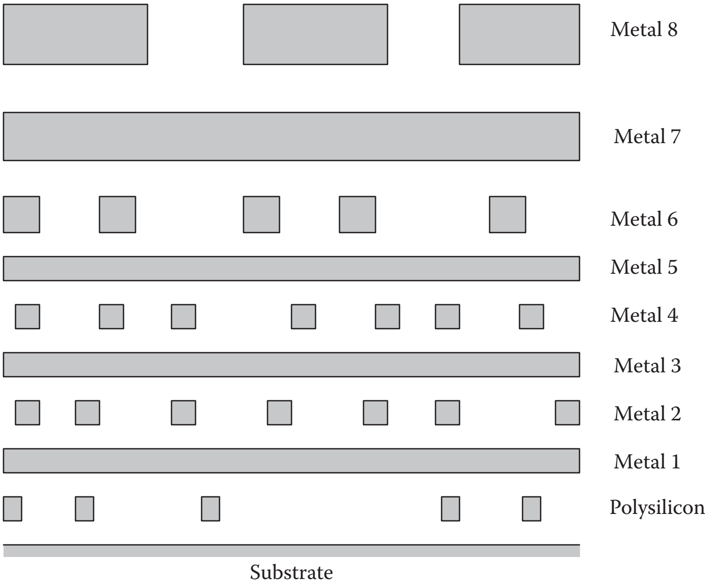


FIGURE 7.10
Orthogonal interconnects on adjacent levels to minimize the crosstalk between levels.

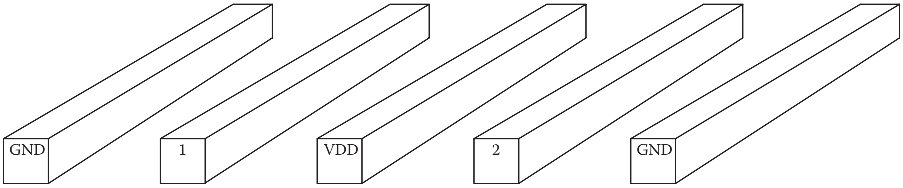


FIGURE 7.11
Power rails inserted between signal lines to reduce the effects of crosstalk.

7.7. Polysilicon Interconnect

Polysilicon is highly resistive compared with metal, even when heavily doped n-type. Thus, long polysilicon interconnects introduce significant delays. However, there are situations in which the use of long polysilicon runs increase the layout efficiency. An example of this is the row lines in a digital memory. In such cases, it is necessary to reduce the interconnect RC delays by using low-resistivity straps (either metal or silicide) or by using repeaters or some combination of both.

Example 7.12 Elmore Delay in Polysilicon Interconnect

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 SiO_2 ($\varepsilon_r = 3.9$).

Solution: The capacitance to the substrate per unit length is

$$c = \varepsilon_{DI}\left[\frac{\left(w-\frac{h}{2}\right)}{t}+\frac{2\pi}{\ln\left(1+\frac{2t}{h}+\sqrt{\frac{2t}{h}\left(\frac{2t}{h}+2\right)}\right)}\right]$$

$$= (3.9)\left(8.85\times10^{-14}F/cm\right)\left[\frac{0.125\times10^{-4}cm}{0.5\times10^{-4}cm}+\frac{2\pi}{\ln\left(1+4+\sqrt{4(4+2)}\right)}\right].$$

$$= 1.03\times10^{-12}F/cm = 1.03pF/cm$$

The resistance per unit length is

$$r = \frac{\rho}{wh} = \frac{(0.01\Omega cm)}{\left(0.25\times10^{-4}cm\right)\left(0.25\times10^{-4}cm\right)} = 16M\Omega/cm.$$

The delay associated with 100 μm of this polysilicon interconnect is

$$t_P \approx \ln(2)\frac{rcl^2}{2} = \ln(2)\frac{\left(16\times10^6\Omega / cm\right)\left(1.03\times10^{-12}F / cm\right)\left(10^{-2}cm\right)^2}{2} = 570ps\,.$$

This is many times the on-chip propagation delay for a modern CMOS logic circuit.

The long RC delays associated with polysilicon interconnect can be dealt with by strapping or the use of repeaters. Strapping involves the implementation of a parallel interconnect with lower resistance per unit length. One approach is the creation of 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 the insertion of buffers at periodic intervals as shown in Figure 7.12.

Example 7.12 Repeaters for Polysilicon Interconnect

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 SiO_2 ($\varepsilon_r = 3.9$), by the use of nine repeaters.

Solution: The capacitance to the substrate per unit length is

$$c = 1.03pF / cm\,.$$

The resistance per unit length is

$$r = 16M\Omega / cm\,.$$

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

$$t_P\left(\text{each segment}\right) \approx \ln(2)\frac{\left(16\times10^6\Omega/cm\right)\left(1.03\times10^{-12}F/cm\right)\left(10^{-3}cm\right)^2}{2} = 5.7ps.$$

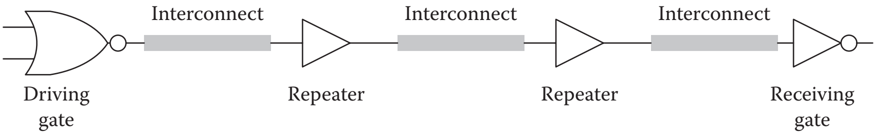


FIGURE 7.12
The use of repeaters to reduce interconnect delays.

If the repeater propagation delays can be neglected, the total delay is

$$t_P \approx 10t_P\left(\text{each segment}\right) = 57ps,$$

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 1/5 reduction in the overall delay, a significant improvement.

7.8 SPICE Demonstrations

For the purpose of illustration, simulations were performed using Cadence Capture CIS 10.1.0 PSpice (Cadence Design Systems). The level 1 MOS transistor model parameters given in Tables 7.4 and 7.5 were used unless otherwise noted. The process transconductance parameters were calculated assuming an oxide thickness of 9 nm. For n-MOSFETS,

$$KP = \frac{(3.9)\left(8.85\times10^{-14}F/cm\right)\left(580cm^2V^{-1}s^{-1}\right)}{9\times10^{-7}cm} = 222\mu A/V^2, \qquad (7.23)$$

and for p-MOSFETS,

$$KP = \frac{(3.9)\left(8.85\times10^{-14}F/cm\right)\left(230cm^2V^{-1}s^{-1}\right)}{9\times10^{-7}cm} = 88\mu A/V^2, \qquad (7.24)$$

TABLE 7.4

n-MOS Level 1 SPICE Parameters

Parameter	Value	Units
KP	222u	A/V^2
VTO	0.5	V
GAMMA	0.15	V$^{1/2}$
PHI	0.7	V
LAMBDA	0.05	
TOX	9n	m
NSUB	1E16	cm^{-3}
UO	580	cm^2/Vs
CGSO	0.38n	F/m
CGDO	0.38n	F/m

TABLE 7.5

p-MOS Level 1 SPICE Parameters

Parameter	Value	Units
KP	88u	A/V^2
VTO	0.5	V
GAMMA	0.15	V$^{1/2}$
PHI	0.7	V
LAMBDA	0.05	
TOX	9n	m
NSUB	1E16	cm^{-3}
UO	230	cm^2/Vs
CGSO	0.38n	F/m
CGDO	0.38n	F/m

The overlap capacitances per unit gate width were determined with the assumption that $L_{OV} = 0.1\mu m$:

$$CGSO = \frac{(3.9)\left(8.85\times 10^{-14} F/cm\right)\left(0.1\times 10^{-4} cm\right)}{9\times 10^{-7} cm} \qquad (7.25)$$

$$= 3.8pF/cm = 0.38nF/m$$

and

$$CGDO = \frac{(3.9)\left(8.85\times 10^{-14} F/cm\right)\left(0.1\times 10^{-4} cm\right)}{9\times 10^{-7} cm} \qquad (7.26)$$

$$= 3.8pF/cm = 0.38nF/m.$$

The body effect coefficient was calculated from

$$GAMMA = \frac{\sqrt{2q\varepsilon_{Si}N_a}}{C_{ox}}$$

$$= \frac{\sqrt{2\left(1.602\times 10^{-19} C\right)(11.9)\left(8.85\times 10^{-14} F/cm\right)\left(10^{16} cm^{-3}\right)}}{(3.9)\left(8.85\times 10^{-14} F/cm\right)/9\times 10^{-7} cm}$$

$$\approx 0.15V^{1/2}.$$

SPICE Example 7.1 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

$$c = 1.03pF/cm$$

and

$$r = 16M\Omega / cm \cdot$$

For a 100 μm length of this interconnect, the expected propagation delay is

$$t_P \approx \ln(2)\frac{rcl^2}{2} = \ln(2)\frac{\left(16\times10^6\Omega / cm\right)\left(1.03\times10^{-12}F / cm\right)\left(10^{-2}cm\right)^2}{2} = 570ps.$$

The 100 μm interconnect may be modeled in SPICE using five RC sections as shown in Figure 7.13.

As can be seen in Figure 7.14, the simulated propagation delay for 100 μm of polysilicon is 730 ps, less than 30% more than the calculated delay. Therefore, the use of a finite number of sections provides fair accuracy.

SPICE Example 7.2 Branched RC Lines

Consider branching interconnect as modeled in Figure 7.15. The propagation delays for nodes 1 and 5 with respect to the driving node may be estimated as

$$\begin{aligned} t_{P1} &\approx \ln(2)\sum_{k=1}^{N} C_k R_{1k} \\ &= \ln(2)\left[C_1R_1 + C_2R_1 + C_3R_1 + C_4R_1 + C_5R_1\right] \qquad (7.27) \\ &= \ln(2)\left[\left(C_1 + C_2 + C_3 + C_4 + C_5\right)R_1\right] \\ &= \ln(2)\left[\left(540fF\right)\left(10\Omega\right)\right] = 3.7ps, \end{aligned}$$

and

$$\begin{aligned} t_{P5} &\approx \ln(2)\sum_{k=1}^{N} C_k R_{1k} \\ &= \ln(2)\left[C_1R_1 + C_2R_1 + C_3\left(R_1 + R_3\right) + C_4\left(R_1 + R_3\right) + C_5\left(R_1 + R_3 + R_5\right)\right] \qquad (7.28) \\ &= \ln(2)\left[\left(C_1 + C_2 + C_3 + C_4 + C_5\right)R_1 + \left(C_3 + C_4 + C_5\right)R_3 + C_5R_5\right] \\ &= \ln(2)\left[\left(540fF\right)\left(10\Omega\right) + \left(360fF\right)\left(24\Omega\right) + \left(120fF\right)\left(40\Omega\right)\right] = 13.1ps. \end{aligned}$$

The delays may be determined in PSpice using a transient simulation and the circuit of Figure 7.16. The pulse source was set up with the following parameters: V1 = 0, V2 = 5 V, TR = 0, TF = 0, PW = 100 ps, and PER = 200 ps.

The results of the simulation appear in Figure 7.17. 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

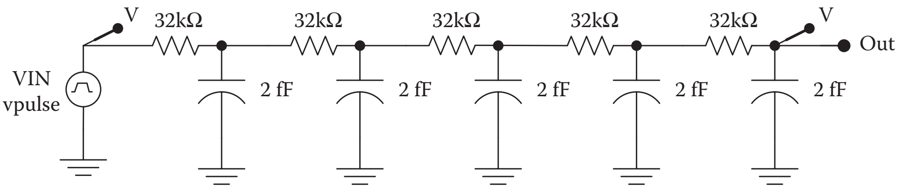


FIGURE 7.13
Circuit for SPICE simulation of a 100 μm polysilicon interconnect with $r = 16M\Omega / cm$ and $c = 1.03pF / cm$, broken up into five identical RC sections.

delay by more than a factor of four (3.7 versus 0.8 ps). Therefore, caution must be used in applying the Elmore approximation to nodes near the driving node.

SPICE Example 7.3 Transmission Line with 1 kΩ Termination

SPICE simulators provide models for both ideal and lossy transmission lines. Here a lossy transmission line element was used, with model parameters C, L, and G, which are the capacitance, inductance, and conductance per unit length, and LEN, which is the length. The lossy transmission line was connected to a pulse source and a terminating resistance of 1 kΩ as shown in Figure 7.18. To model 1 cm of copper interconnect, the transmission model parameters were set up as follows: C = 1.38E-12 F/m, L = 3.2E-9 H/m, G = 0, and LEN = 0.01 m. The pulse

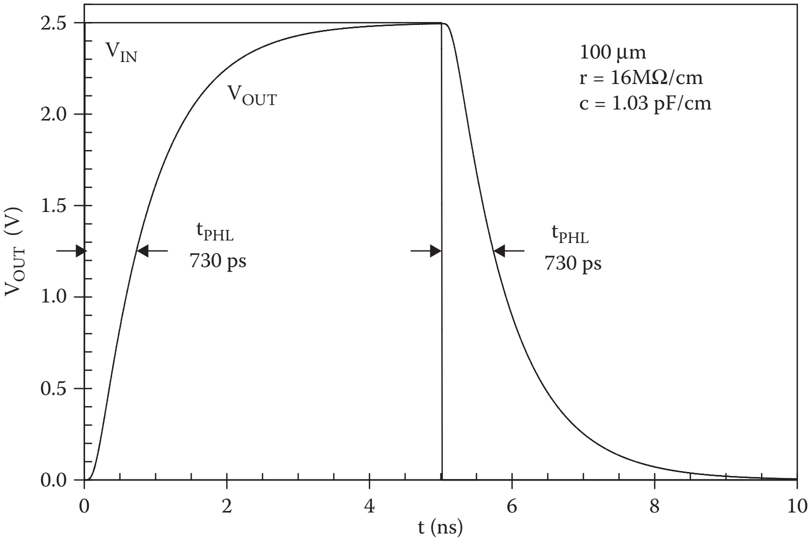


FIGURE 7.14
Transient simulation for a 100 μm polysilicon interconnect modeled by five identical RC sections.

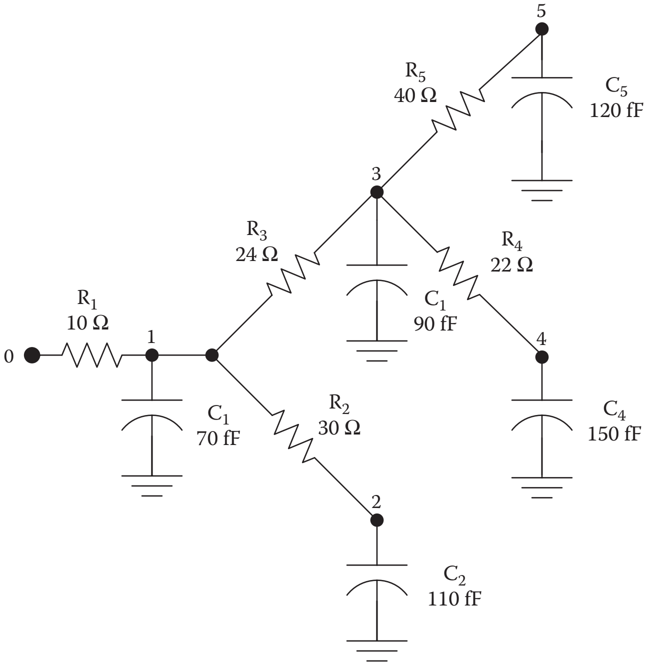


FIGURE 7.15
Branching interconnect modeled by a branching RC network.

source was set up with the following parameters: V1 = 0, V2 = 5 V, TR = 0, TF = 0, PW = 20 ps, and PER = 2000 ps. A source impedance of 10 Ω was included, and, for this configuration, the reflection coefficients are 0.90 at the load and – 0.67 at the source.

The results in Figure 7.19 illustrate the importance of reflections in a transmission line that is not terminated with an impedance-matched load. (In this case, the matched termination would be ~50 Ω.) 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 itself 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, or the time for "out and back" travel down the transmission line, and each successive pulse is of opposite sign because of the negative reflection coefficient at the source.

SPICE Example 7.4 Transmission Line with 50 Ω Termination

If the transmission line from the previous example is terminated with a matched 50 Ω load as in Figure 7.20, very different behavior is observed. As shown in Figure 7.21, 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

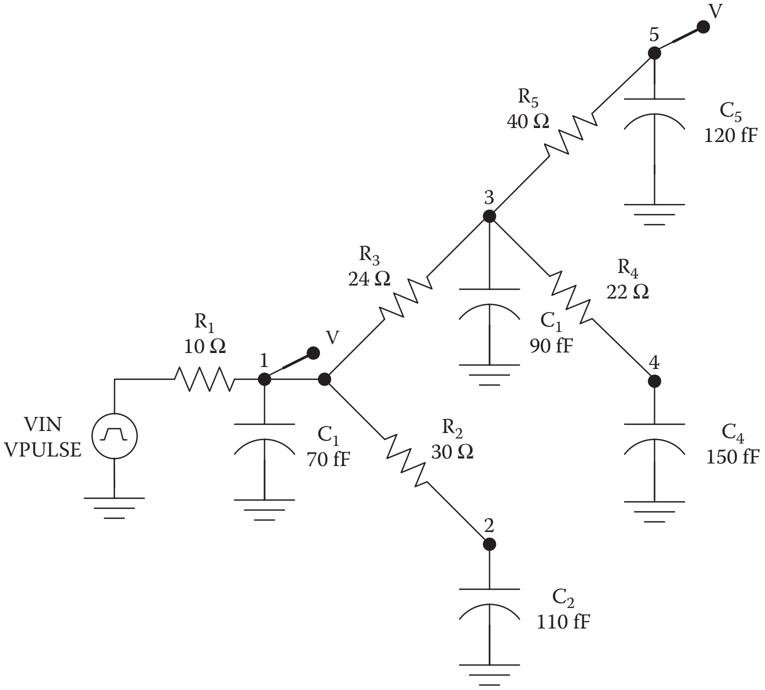


FIGURE 7.16
Circuit model used for the SPICE simulation of the transient response for branching interconnect.

the matched load is zero so that reflected pulses do not travel back and forth in the line, although the source impedance is unmatched to the transmission line. Together, these two examples show that signal reflections may be problematic in interconnects or circuit board traces that act as transmission lines unless impedance-matched terminations are used.

SPICE Example 7.5 Interconnect and Gate Delay Interactions

Consider a situation in which a CMOS inverter drives a similar CMOS inverter through a 100 μm section of polysilicon interconnect as shown in Figure 7.22. The parasitics of the polysilicon interconnect are assumed to be $c = 1.03pF / cm$ and $r = 16M\Omega / cm$, and the interconnect is therefore approximated by five RC sections, each with $C = 2fF$ and $r = 32k\Omega$. The pulse source driving this arrangement has been set up with V1 = 0, V2 = 2.5 V, TR = 0, TF = 0, PW = 10 ns, and PER = 20 ns, and for the symmetric CMOS inverters $V_{DD} = 2.5V$. The overall delay from the pulse source input to the output of the second inverter is 2.0 ns, and most of this delay is attributable to the interconnect. The interconnect delay in this circuit is considerably more than what it would be for the

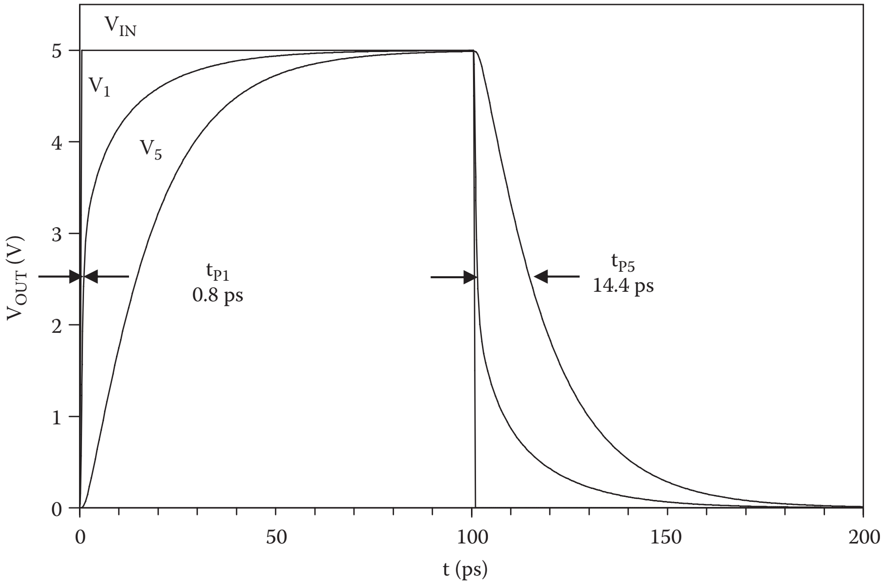


FIGURE 7.17
Simulated transient response for the branching interconnect.

interconnect alone (570 ps) attributable to the capacitive loading of the second inverter (Figure 7.23).

SPICE Example 7.6 CMOS Inverters Connected by Transmission Line

Now consider that the inverters from the previous example are connected by a 1 cm length of copper interconnect modeled as a transmission line as shown in Figure 7.24: C = 1.38E-12 F/m, L = 3.2E-9 H/m, G = 0, and LEN = 0.01 m. The pulse source was set up with the following parameters: V1 = 0, V2 = 5 V, TR = 0, TF = 0, PW = 5 ns, and PER = 10 ns. The simulation results in Figure 7.25 are qualitatively similar

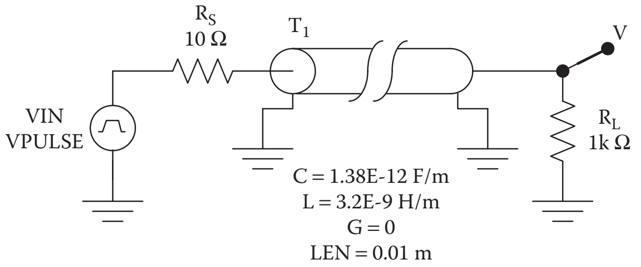


FIGURE 7.18
SPICE circuit for transient simulation of a transmission line with a 1 kΩ termination.

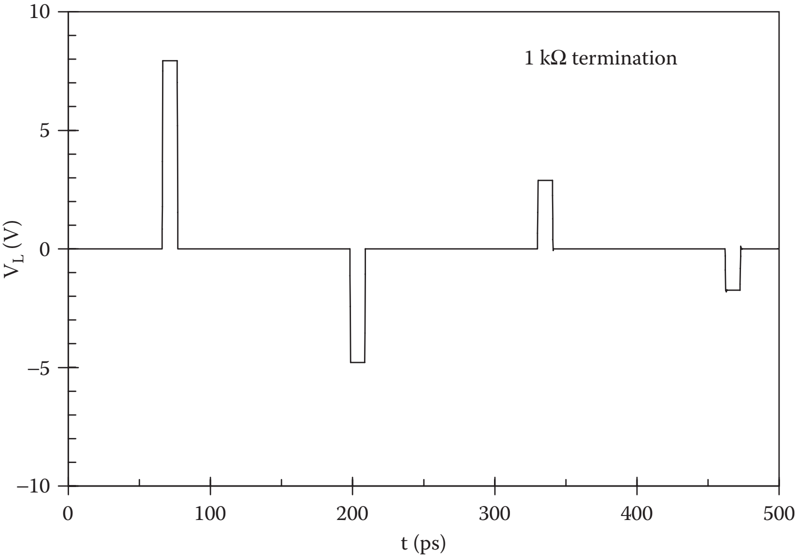


FIGURE 7.19
SPICE transient simulation for a 50 Ω transmission line with a 1 kΩ termination.

to those obtained in the previous example. However, because neither the sending gate nor the receiving gate is impedance matched to the transmission line, reflected signals appear as a series of glitches separated by twice the time of flight for the transmission line. (The time of flight is 66 ps; the glitches are separated by 132 ps.)

7.9 Practical Perspective

For practical perspective articles, see the dynamic website at http://www.engr.uconn.edu/ece/books/ayers.

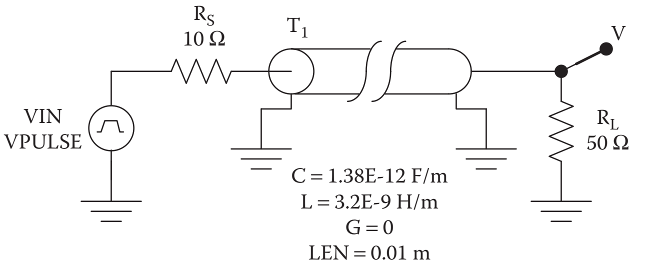


FIGURE 7.20
SPICE circuit for transient simulation of a transmission line with a matched 50 Ω termination.

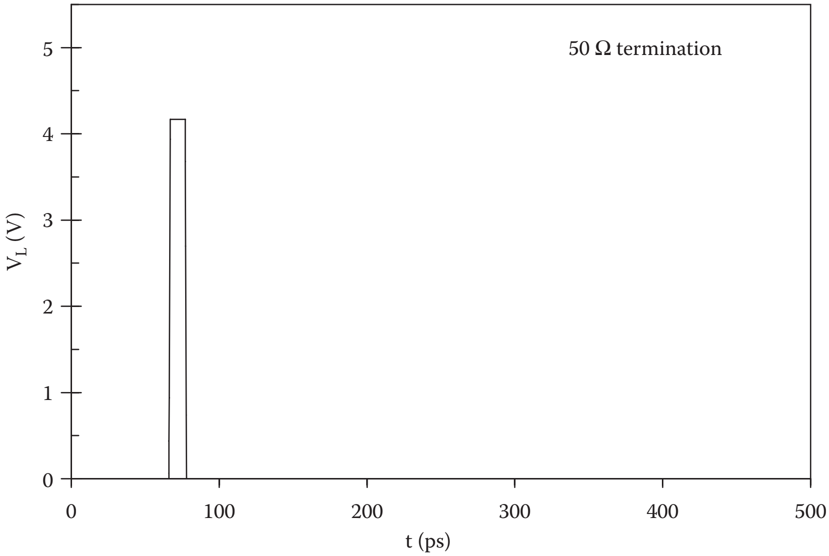


FIGURE 7.21
SPICE transient simulation for a transmission line with a matched 50 Ω termination.

7.10 Summary

VLSI digital integrated circuits use up to several kilometers of interconnections made from copper, aluminum, and polysilicon. These interconnects exhibit parasitic capacitances and resistances that increase circuit delay times as well as the overall power dissipation, making interconnect an important design consideration for integrated circuits.

Simple equations have been presented for the calculation of the parasitic resistances, capacitances, and inductances on a per unit length basis. The capacitance to ground per unit length includes parallel plate capacitance and fringing field components, but typically the fringing field capacitance

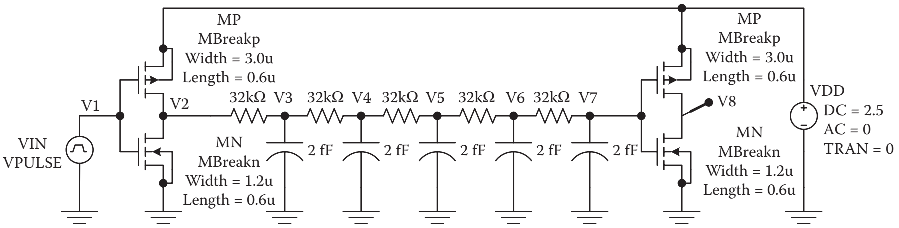


FIGURE 7.22
Symmetric CMOS inverters connected through a 100 μm section of polysilicon interconnect. The interconnect is modeled using five identical RC sections.

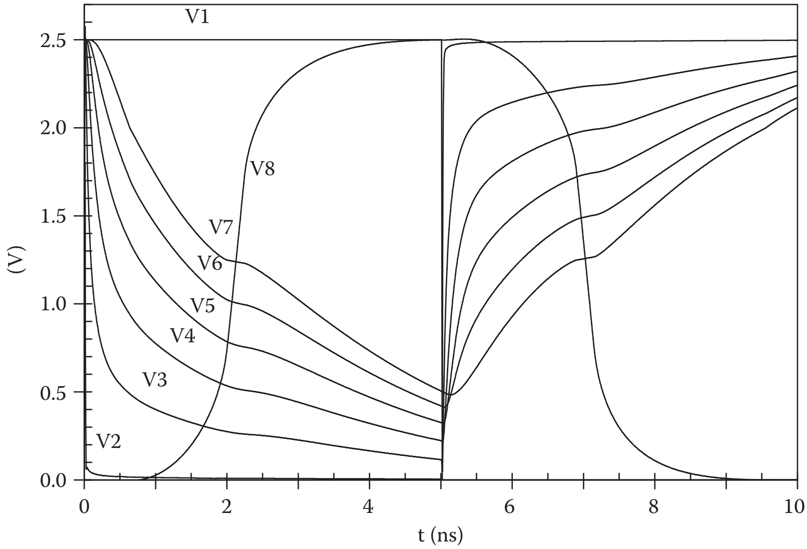


FIGURE 7.23
Transient simulation results for symmetric CMOS inverters connected through a 100 μm length of polysilicon interconnect, modeled as five RC sections.

is dominant. Capacitances between wires are also significant in a multilevel metallization scheme. The inductance per unit length may be estimated from the approximate value of capacitance per unit length if the permeabilities and permittivities of the insulator are known. However, the inductance is significant only in very long interconnects or those providing power distribution.

Short lengths of interconnect may be modeled using simple lumped capacitors, but longer interconnects should be modeled as distributed RC networks. Very long interconnects may behave as transmission lines, so that impedance-matched terminations will be necessary to avoid reflected signals.

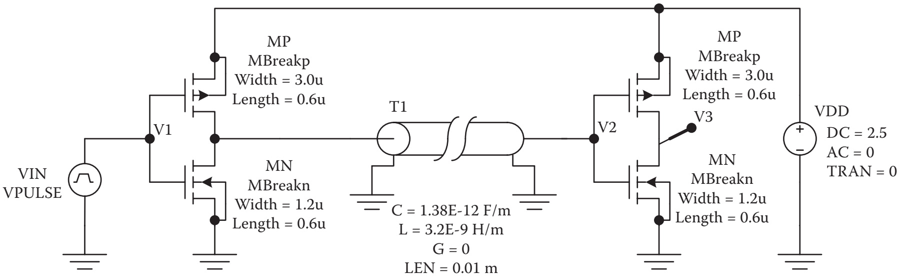


FIGURE 7.24
Symmetric CMOS inverters connected through 1 cm length of copper interconnect modeled as a transmission line with C = 1.38E-12 F/m, L = 3.2E-9 H/m, G = 0, and LEN = 0.01 m.

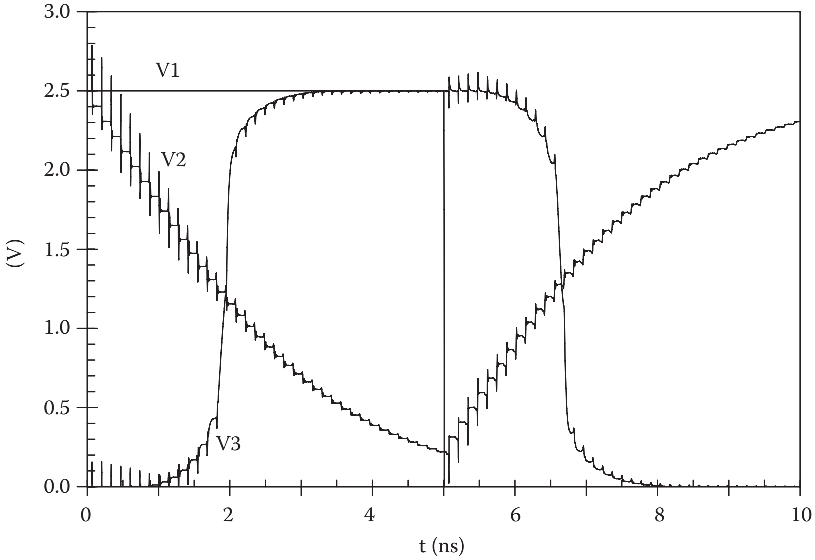


FIGURE 7.25
Transient simulation results for two CMOS inverters connected by a 1 cm transmission line as shown in Figure 7.24. V1 is the input to the sending inverter, V2 is the voltage at the input of the receiving gate, and V3 is the voltage at the output of the receiving gate.

There is a complex interplay between the delays introduced by CMOS gates and interconnect wires, so that SPICE modeling is necessary to predict the overall performance.

7.11 Exercises

E7.1. 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) and n-type polysilicon.

E7.2. Estimate the skin depth at 20 GHz for each of the following interconnect materials: (1) copper, (2) aluminum, and (3) n-type polysilicon.

E7.3. Estimate the capacitance to ground per unit length for a copper interconnect 1.0 μm in height and with widths of 0.5, 1.0, and 2.0 μm. The dielectric is SiO_2, 1.5 μm thick.

E7.4. Estimate the inductance per unit length for the interconnects of Exercise E7.2.

E7.5. Estimate the propagation delay for 100 μm of p-type polysilicon, 0.25 μm wide and 1.0 μm in height.

E7.6. 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 with $K_N = K_P = 0.2mA / V^2$, $V_{DD} = 1.0V$, and $V_{TN} = |V_{TP}| = 0.3V$.

E7.7. For the branching interconnect, modeled by a branching RC tree as depicted in Figure 7.26, estimate propagation delays from the source node to nodes 1–5.

E7.8. Consider aluminum level 1 interconnect, 0.5 μm wide and 1.0 μm high, on 1.5 μm of carbon-doped SiO_2. Estimate the ranges of length for which the following models are appropriate: (1) the lumped capacitance model, (2) the distributed RC model, and (3) the transmission line model. For the CMOS driving gate, $t_P = C_L(1ps / fF)$.

E7.9. Consider the branching interconnect, modeled by a branching RC tree as shown in Figure 7.27. Suppose the driving gate is a CMOS inverter with $K_N = K_P = 0.2mA / V^2$, $V_{DD} = 1.0V$, and $V_{TN} = |V_{TP}| = 0.3V$. Can the resistances be neglected for the determination of the propagation delays from the input of the inverter to the nodes 1–4?

For additional exercise problems, see the dynamic website at http://www.engr.uconn.edu/ece/books/ayers.

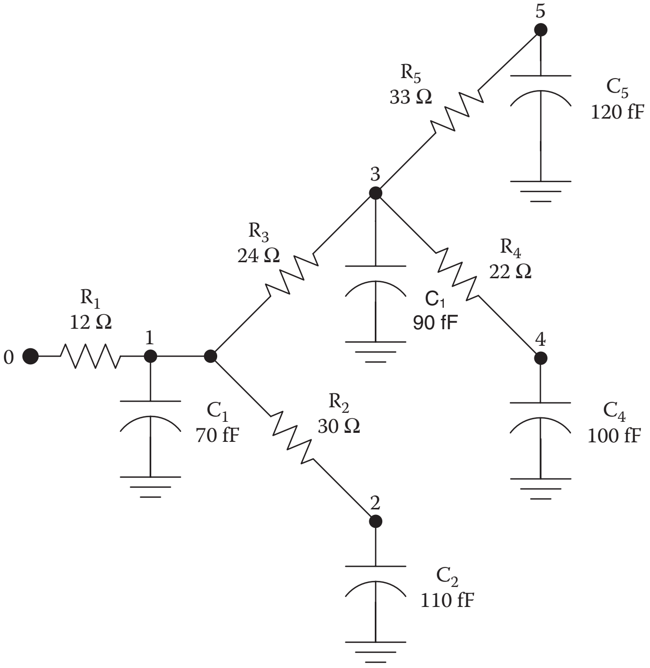


FIGURE 7.26
Branching interconnect for the determination of the delay times (see Exercise E7.6).

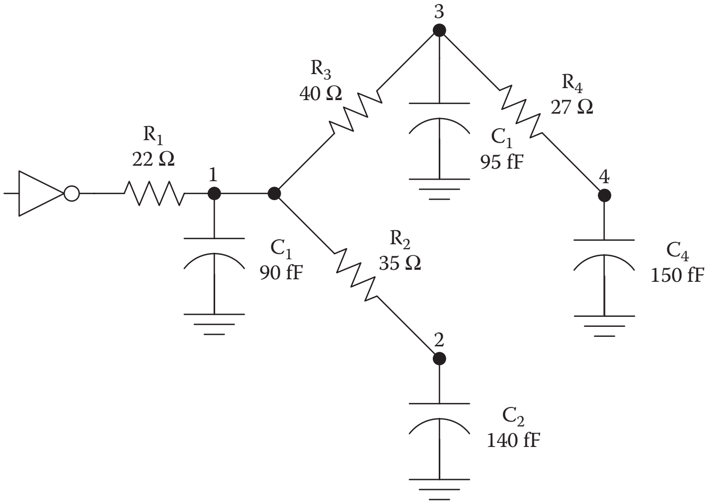


FIGURE 7.27
CMOS inverter driving an RC tree (see Exercise E7.8).

References

1. Meindl, J.D., Beyond Moore's Law: The interconnect era. *Computing Sci. Engr.*, 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 Automat. 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. Yuan, J., and Trick, T. N., A simple formula for the estimation of capacitance of two-dimensional interconnects in VLSI circuits. *IEEE Electron Dev. Lett.*, EDL-3, 391–393, 1982.
9. Wang, S.-Q., Spin-on dielectric films—A general overview. *Proc. 5th Int. Solid-State IC Technol. Conf.*, 961, 1998.

10. Taylor, K.J., Jeng, S.-P., Eissa, M., Gaynor, J., and Nguyen, H., Polymers for high performance interconnects, Abstract Booklet 1997, European Workshop. *Mater. Adv. Metall.*, 59, 1997.
11. Ruelke, H., Streck, C., Hohage, J., Weiher-Telford, S., and Chretrien, O., Manufacturing implementation of low-κ dielectrics for copper damascene technology. *Proc. 2002 IEEE Conf. Adv. Semiconductor Manufact.*, 356, 2002.
12. Lee, P.W., Chi-I Lang; Sugiarto, D., Li-Qun Xia, 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.
13. Mosig, K., Jacobs, T., Kofron, P., Daniels, M., Brennan, K., Gonzales, A., Augur, R., Wetzel, J., Havemann, R., Shiota, A., Single and dual damascene integration of a spin-on porous ultra low-κ material. *Proc. 2001 IEEE Interconnect Technol. Conf.*, 292, 2001.
14. Kawakami, N., Fukumoto, Y., Kinoshita, T., Suzuki, K., and Inoue, K.-I., A super low-κ (κ = 1.1) silica aerogel film using supercritical drying technique. Proc. *IEEE Interconnect Technol. Conf.*, 143, 2000.
15. 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 κ (κ = 2.8) interconnect. *1999 IEEE Int. Conf. Interconnect Technol.*, 181, 1999.
16. 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.
17. 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.
18. Davis, J.A., and Meindl, J.D., Compact distributed RLC models for multilevel interconnect networks, Digest of Technical Papers 1999. *Symp. VLSI Technol.*, 165, 1999.
19. 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 Dev*., 47, 2078, 2000.
20. 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 Dev.*, 50, 1081, 2003.
21. 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 Dev*., 50, 1094, 2003.
22. Elmore, E., The transient response of damped linear networks with particular regard to wideband amplifiers. *J. Appl. Physiol.*, 55, 1948.
23. Yamakoshi, K., and Ino, M., Generalised Elmore delay expression for distributed RC tree networks. *Electronics Lett.*, 29, 617, 1993.
24. 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.
25. Adler, V., and Friedman, E., Uniform repeater insertion in RC trees. *IEEE Trans. Circ. Sys. I Fund. Theory App.,* 47, 1, 2000.

8

Dynamic CMOS

8.1 Introduction

Dynamic, or clocked, CMOS gates achieve greater speed and allow higher packing densities than the static CMOS circuits discussed so far [1]. The improved packing density comes from the fact that most logic functions can be implemented using fewer transistors in a dynamic logic circuit. Reduced dissipation is associated with lower input capacitances, because each input is connected to a single n-MOS transistor. The basic principles underlying the operation of dynamic CMOS can be appreciated by examination of the dynamic CMOS inverter shown in Figure 8.1. This circuit has a periodic clock signal CLK as well as a logic input IN.

During the precharge phase, the clock signal goes low; this causes the precharge transistor MPPRE to turn on and bring the output node to V_{DD}. The time required for this output node to charge depends on the capacitance loading the output node, C_Y, as well as the on current of the precharge transistor. Once charged, the capacitance C_Y will retain a voltage close to

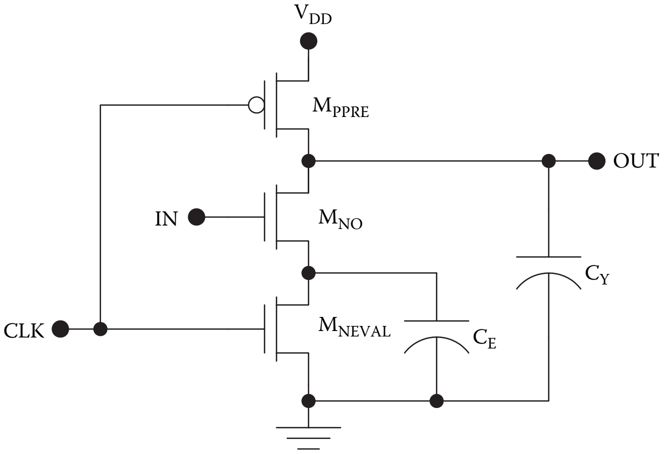


FIGURE 8.1
Dynamic CMOS inverter circuit.

V_{DD} even after the precharge transistor is switched off. (The time this charge will be retained depends on the off currents in the n-MOS devices M_{NO} and M_{NEVAL}.)

During the evaluation phase, the clock signal goes high, turning on M_{NEVAL}. This allows the output node to discharge if but only if the transistor M_{NO} is also on. Therefore, the output voltage read at the end of the evaluation phase will be the inversion of the input signal. The time required for the output to discharge depends on the output node capacitance C_Y and the widths of the two n-MOS transistors.

The basic operation of the dynamic CMOS inverter is illustrated in the timing diagram of Figure 8.2. Here the rise and fall times for the output signal are roughly equal. When this is not true, the worst case of t_R , t_F limits the maximum clock frequency. There is a small glitch in the output voltage when the input transitions from low to high, and this is attributable to charge sharing between the output node capacitance and the capacitance loading the node where M_{NO} and M_{NEVAL} are joined.

In the dynamic CMOS inverter of Figure 8.1, the output node is referred to as a *soft node*. This is because a high output voltage exists only by virtue of the charge stored on C_Y, and is not accompanied by the active drive of a p-MOS pull-up network. The voltage on such a soft node may be easily disturbed by a single event upset caused by cosmic radiation or an alpha particle [2].

Whereas each input to a static CMOS gate is connected to one n-MOS transistor and one p-MOS transistor, the input to the dynamic CMOS gate is applied to a single transistor. Therefore, the driving gate in the previous stage will experience a greatly reduced load capacitance and higher speed operation is possible. This is a key advantage of dynamic CMOS and, along with the higher packing density, is the motivation for using dynamic CMOS in high-performance VLSI circuits.

The main disadvantage of dynamic CMOS is the need for a clock signal with a minimum frequency for correct operation. The leakage currents in the off transistors cause the voltages on the soft nodes to gradually deteriorate.

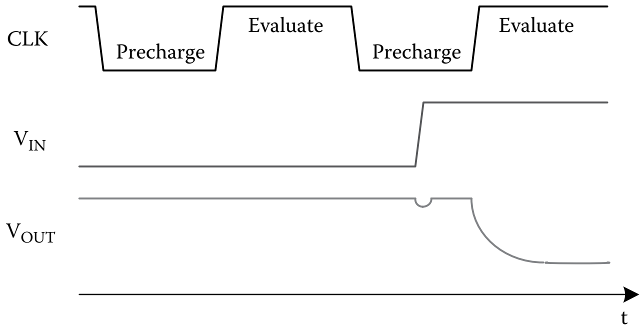


FIGURE 8.2
Timing diagram for dynamic CMOS inverter circuit.

Therefore, periodic refreshing is necessary even when the inputs are not changing. This contrasts with static CMOS, which can be slowed arbitrarily or even stopped. The requirement of routing the clock signal to each gate circuit is a disadvantage in itself, because the clock represents a complex routing problem in a large-area VLSI circuit attributable to *clock skew* (point-to-point phase differences in the clock resulting from unequal propagation delays) [3].

There are several issues that impact the power dissipation of dynamic CMOS. The total switched capacitance may be reduced because of the lower input capacitance. However, this benefit may be offset by the increased switching capacitance associated with the clock signal and also the need to switch all circuits at a minimum frequency. Moreover, correct operation of dynamic CMOS circuits at a particular frequency requires a minimum ratio of I_{ON}/I_{OFF} for the transistors. This may dictate a higher supply voltage V_{DD} than would be used in static CMOS gates using similar transistors, and the switching dissipation increases with the square of V_{DD}. In the balance, the dissipation for dynamic circuits may actually be higher than that for static CMOS gates.

In the following sections, some of the key considerations for dynamic CMOS circuits will be examined in more detail.

8.2 Rise Time

A basic speed limitation for a dynamic CMOS inverter is the rise time for the soft output node. To estimate this, we will assume the clock signal has zero fall time, that the initial voltage at the soft output node is zero, and that this node is loaded with a lumped capacitance C_Y as shown in Figure 8.3. If the CLK signal falls abruptly, the evaluation device M_{NEVAL} will turn off abruptly, and we do not need to consider the conduction of M_{NO} for a first-order calculation, even if the input signal is high. The charge-up time is the sum of two components $t_R = t_{R1} + t_{R2}$, where t_{R1} and t_{R2} represent the time intervals for saturated and linear operation of the p-MOS precharge transistor, respectively. If the output voltage is initially zero and CLK makes an abrupt transition, the precharge transistor will be saturated until V_{OUT} increases to $-V_{TP}$. During this time, interval a constant current charges the soft node capacitance so that

$$t_{R1} = \int_0^{-V_{TP}} \frac{C_Y dV_{OUT}}{K_P \left(V_{DD} + V_{TP}\right)^2 / 2} = -\frac{-2V_{TP}C_Y}{K_P \left(V_{DD} + V_{TP}\right)^2} . \tag{8.1}$$

After V_{OUT} reaches $-V_{TP}$, the p-MOS device operates in the linear mode until the end of the charge-up time. Using the usual definition for the rise time,

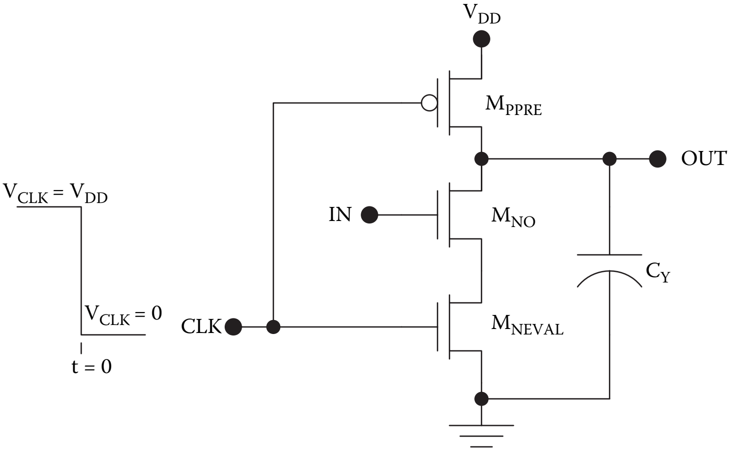


FIGURE 8.3
Dynamic CMOS inverter for the consideration of the output rise time.

which corresponds to $V_{OUT} = 0.9V_{DD}$ (the 90% point), this time interval is given by

$$\begin{aligned} t_{R2} &= \int_{-V_{TP}}^{0.9V_{DD}} \frac{C_Y dV_{OUT}}{K_P\left[(V_{DD}+V_{TP})(V_{DD}-V_{OUT})-(V_{DD}-V_{OUT})^2/2\right]} \\ &= \frac{C_Y}{K_P(V_{DD}+V_{TP})}\ln\left(\frac{19V_{DD}+20V_{TP}}{V_{DD}}\right). \end{aligned} \tag{8.2}$$

The rise time is the sum $t_{R1} + t_{R2}$, so that

$$t_R = \frac{C_Y}{K_P(V_{DD}+V_{TP})}\left[\frac{-2V_{TP}}{(V_{DD}+V_{TP})}+\ln\left(\frac{19V_{DD}+20V_{TP}}{V_{DD}}\right)\right]. \tag{8.3}$$

This is the same as the rise time for a static CMOS inverter with a load capacitance C_Y.

8.3 Fall Time

The fall time t_F for a dynamic CMOS inverter may be estimated for the situation depicted in Figure 8.4 with the assumptions that (1) the CLK signal makes an abrupt low-to-high transition, (2) the initial output voltage is V_{DD},

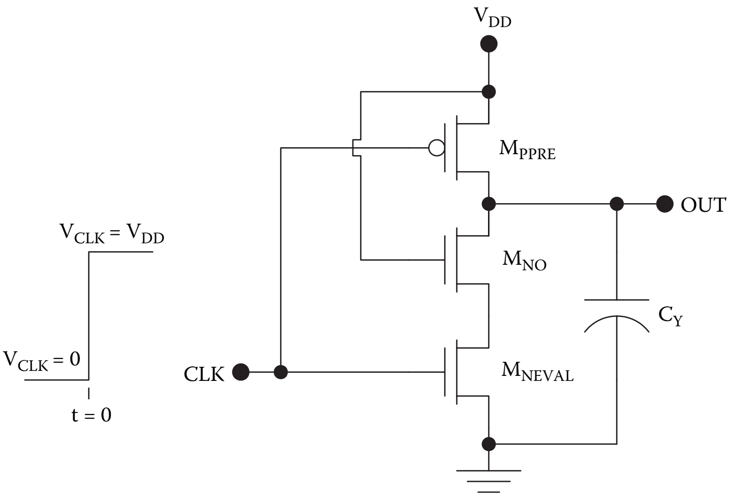


FIGURE 8.4
Dynamic CMOS inverter for the consideration of the output charge-down time.

(3) the soft output node is loaded by a capacitance C_Y, and (4) the capacitance loading the drain of the evaluation transistor may be neglected.

If the input voltage is tied to VDD, the output node will make a high-to-low transition after the CLK goes high. The fall time is the sum of two components $t_F = t_{F1} + t_{F2}$, where t_{F1} and t_{F2} represent the time intervals for saturated and linear operation of the n-MOS transistor, respectively. The n-MOS transistor operates in the saturation region until V_{OUT} drops to $V_{DD} - V_{TN}$, and the length of this time interval is

$$t_{F1} = \frac{2V_{TN}C_Y}{K_N\left(V_{DD} - V_{TN}\right)^2}\,. \tag{8.4}$$

After V_{OUT} drops below $V_{DD} - V_{TN}$, the n-MOS operates in the linear mode until the end of the fall time, which corresponds to $V_{OUT} = V_{DD}/10$ (the 10% point). The length of this time interval is

$$\begin{aligned} t_{F2} &= -C_Y \int_{V_{DD}-V_{TN}}^{V_{DD}/10} \frac{dV_{OUT}}{K_N\left[\left(V_{DD} - V_{TN}\right)V_{OUT} - V_{OUT}^2/2\right]} \\ &= -\frac{C_Y}{K_N\left(V_{DD} - V_{TN}\right)}\left[ln\left(\frac{V_{OUT}}{V_{OUT} - 2\left(V_{DD} - V_{TN}\right)}\right)\right]_{V_{DD}-V_{TN}}^{V_{DD}/10} \\ &= \frac{C_Y}{K_N\left(V_{DD} - V_{TN}\right)}\ln\left(\frac{19V_{DD} - 20V_{TN}}{V_{DD}}\right). \end{aligned} \tag{8.5}$$

The fall time is given by the sum $t_{F1} + t_{F2}$, so that

$$t_F = \frac{C_Y}{K_N\left(V_{DD} - V_{TN}\right)}\left[\frac{2V_{TN}}{\left(V_{DD} - V_{TN}\right)} + \ln\left(\frac{19V_{DD} - 20V_{TN}}{V_{DD}}\right)\right]. \tag{8.6}$$

This is the same as the fall time for a static CMOS inverter with a load capacitance C_Y.

Example 8.1 Rise and Fall Times for a Dynamic CMOS Inverter

Estimate t_R and t_F for the dynamic CMOS inverter depicted in Figure 8.5 with V_{DD} = 2.5 V, $V_{TN} = |V_{TP}|$ = 0.5 V, and t_{OX} = 9 nm, assuming that the CLK signal makes abrupt transitions.

Solution: The device transconductance parameters are

$$K_P = \frac{W_P}{L_P}\frac{\mu_p\varepsilon_{OX}}{t_{OX}} = \left(\frac{1.2\mu m}{0.6\mu m}\right)\frac{\left(230cm^2/Vs\right)(3.9)\left(8.85\times10^{-14}F/cm\right)}{9\times10^{-7}cm}$$

$$= 176\mu A/V^2$$

and

$$K_N = \frac{W_N}{L_N}\frac{\mu_n\varepsilon_{OX}}{t_{OX}} = \left(\frac{1.2\mu m}{0.6\mu m}\right)\frac{\left(580cm^2/Vs\right)(3.9)\left(8.85\times10^{-14}F/cm\right)}{9\times10^{-7}cm}$$

$$= 445\mu A/V^2.$$

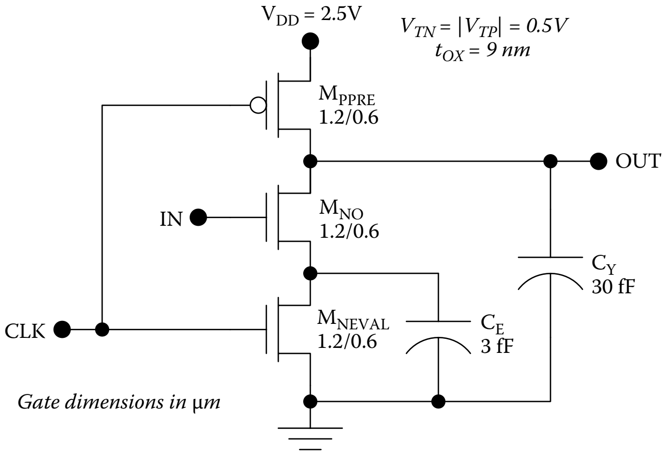


FIGURE 8.5
Example dynamic inverter for the calculation of the rise and fall times.

The rise time is

$$t_R = \frac{C_Y}{K_P\left(V_{DD}+V_{TP}\right)}\left[\frac{-2V_{TP}}{\left(V_{DD}+V_{TP}\right)}+\ln\left(\frac{19V_{DD}+20V_{TP}}{V_{DD}}\right)\right]$$

$$= \frac{30\times10^{-15}F}{\left(176\times10^{-6}A/V^2\right)\left(2.5V-0.5V\right)}\left[\frac{2\left(0.5V\right)}{\left(2.5V-0.5V\right)}\right.$$

$$\left.+\ln\left(\frac{19\left(2.5V\right)-20\left(0.5V\right)}{2.5V}\right)\right]$$

$$= 270ps$$

and the fall time is

$$t_F = \frac{C_Y}{K_N\left(V_{DD}-V_{TN}\right)}\left[\frac{2V_{TN}}{\left(V_{DD}-V_{TN}\right)}+\ln\left(\frac{19V_{DD}-20V_{TN}}{V_{DD}}\right)\right]$$

$$= \frac{30\times10^{-15}F}{\left(446\times10^{-6}A/V^2\right)\left(2.5V-0.5V\right)}\left[\frac{2\left(0.5V\right)}{\left(2.5V-0.5V\right)}\right.$$

$$\left.+\ln\left(\frac{19\left(2.5V\right)-20\left(0.5V\right)}{2.5V}\right)\right]$$

$$= 110ps.$$

The clock frequency must be chosen so that one-half of the period is longer than the worst case of these two delays.

8.4 Charge Sharing

The output voltage of a dynamic CMOS circuit can be affected by changes in the input voltage(s) even during the precharge phase (CLK low). This is a consequence of charge sharing [4] and may be understood with the benefit of Figure 8.6. Here the CLK signal is low so M_{PPRE} is linear but M_{NEVAL} is cutoff. If C_Y has already been charged to V_{DD} but the input voltage makes a low-to-high transition, M_{NO} will turn on and connect the capacitance C_E to the output node. This will cause a momentary drop in the output voltage because charge from C_Y will be shared with the smaller capacitance C_E. If the voltage on the capacitor C_E is initially zero, the current flowing in M_{NO} will greatly

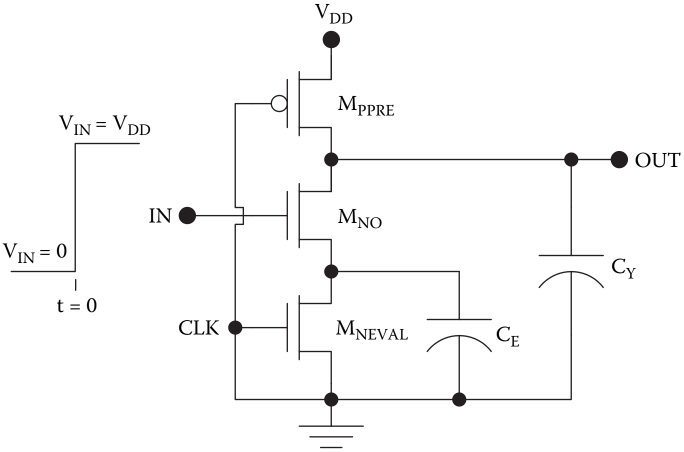


FIGURE 8.6
Charge sharing in a dynamic CMOS inverter.

exceed that in M_{PPRE}. Therefore, we can estimate the maximum excursion of the output voltage by neglecting the drain current in the precharge transistor. Hence, the minimum value of V_{OUT} will be

$$V_{OUT,\min} = \frac{V_{DD}C_Y}{C_E + C_Y}. \tag{8.7}$$

For example, if C_E is 10% of C_Y, then the output voltage will drop by about 10% before being pulled back up to V_{DD}.

Charge sharing is a critical consideration for dynamic CMOS circuit design because, for low fan-out situations, the capacitances of internal soft nodes become comparable with the output node capacitance. On the other hand, charge sharing is of no consequence in static CMOS circuits because the output is actively pulled up or down at every point in time.

8.5 Charge Retention

The output voltage of a dynamic CMOS gate deteriorates slowly with time because of the low leakage currents associated with CMOS circuitry. Eventually the signal must be refreshed, and this necessitates that the clock run at some minimum frequency to maintain signal integrity. This is a disadvantage compared with static CMOS, which can be slowed down arbitrarily or even stopped.

The loss of charge from a soft node in a dynamic circuit is governed by subthreshold leakage currents in the MOSFETs and sometimes the reverse leakage currents of their p-n junctions. Consider the dynamic CMOS inverter of Figure 8.7. The soft output node discharges when CLK is high and the precharge transistor is cutoff. If the input voltage is low, M_{NO} is cutoff as well. There will be p-n junction reverse currents I_{pnPRE}, associated with the drain junction of M_{PPRE}, and I_{pnO}, associated with the drain junction of M_{NO}, as well as the subthreshold current I_{subO} flowing in M_{NO}. Therefore, the leakage current discharging the capacitor is

$$I_{leak} = I_{subO} + I_{pnO} - I_{pnPRE}\,. \tag{8.8}$$

If the drain p-n junction leakage currents of M_{NO} and M_{PPRE} are roughly equal, $I_{pnO} \approx I_{pnPRE}$, the subthreshold current is dominant:

$$I_{leak} \approx K_N\,(m-1)\left(\frac{kT}{q}\right)^2 \exp\left(-\frac{qV_{TN}}{mkT}\right) = C_Y\,\frac{dV_Y}{dt}\,. \tag{8.9}$$

The retention time is the time after which the soft node voltage will have degraded to the midpoint or switching threshold, which can be estimated as 50% of V_{DD}. Thus,

$$t_{ret} \approx \frac{C_Y V_{DD}}{2K_N\,(m-1)(kT/q)^2 \exp(-qV_{TN}/mkT)}\,. \tag{8.10}$$

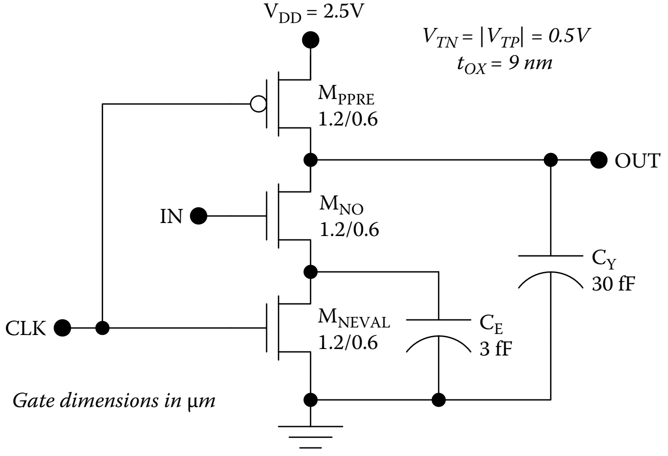


FIGURE 8.7
Dynamic CMOS inverter for the consideration of the retention time.

This simple model can be extended to other circuits and situations as well. Generally, the subthreshold currents will be dominant in circuits with submicron transistors.

Example 8.2 Charge Retention in a Dynamic CMOS Circuit

Estimate the charge retention time t_{ret} for the dynamic CMOS inverter of Figure 8.8, assuming that m = 1.6 for subthreshold operation of the MOSFETs.

Solution: The device transconductance parameter for M_{NO} is

$$K_N = \frac{W_N}{L_N}\frac{\mu_n \varepsilon_{OX}}{t_{OX}} = \left(\frac{1.2\mu m}{0.6\mu m}\right)\frac{(580cm^2/Vs)(3.9)(8.85\times10^{-14}F/cm)}{9\times10^{-7}cm} = 440\mu A/V^2.$$

The retention time is

$$t_{ret} \approx \frac{C_Y V_{DD}}{2K_N(m-1)(kT/q)^2\exp(-qV_{TN}/mkT)}$$

$$= \frac{(30\times10^{-15}F)(2.5V)}{2(440\times10^{-6}A/V^2)(1.6-1)(0.026V)^2\exp(-0.5V/0.040V)}$$

$$= 35ms.$$

The retention time is usually on the order of milliseconds in most dynamic circuits, including dynamic memory cells considered in Chapter 11.

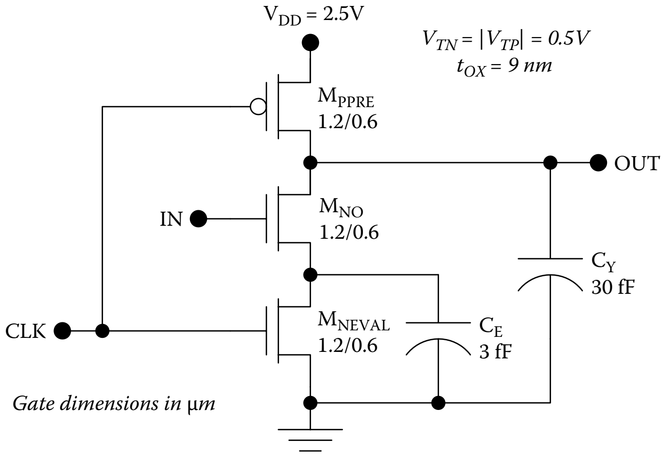


FIGURE 8.8
Example dynamic inverter for the calculation of the retention time.

8.6 Logic Design

Other logic functions can be realized in dynamic CMOS using a single precharge transistor and a single evaluation transistor. The general structure of a dynamic CMOS gate with an n-MOS pull-down network is shown in Figure 8.9. 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 a static CMOS gate. Therefore, with N inputs the dynamic CMOS gate requires N + 2 transistors compared with 2N transistors for a static CMOS gate. This difference, along with the scaling requirements imposed on the p-MOSFETs in static CMOS, account for the higher packing densities in dynamic CMOS logic.

Figure 8.10 shows a three-input dynamic CMOS NOR gate, and Figure 8.11 shows a two-input dynamic CMOS NAND gate. The precharge and evaluate transistors may be 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 so that it is desirable to scale the widths of these devices by a factor of two compared with the reference inverter. More complicated logic functions may be realized also, as long as the transistors in the pull-down network are scaled appropriately.

Although one approach to sizing transistors in CMOS transistors has been described here, other schemes exist as well. For example, the widths of the series-connected n-MOS transistors in the NAND gate may be tapered, with

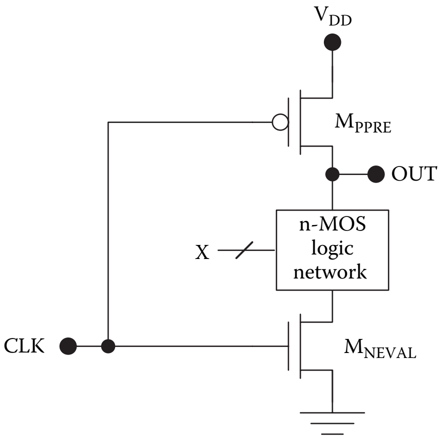


FIGURE 8.9
General dynamic CMOS logic circuit with an n-MOS pull-down network.

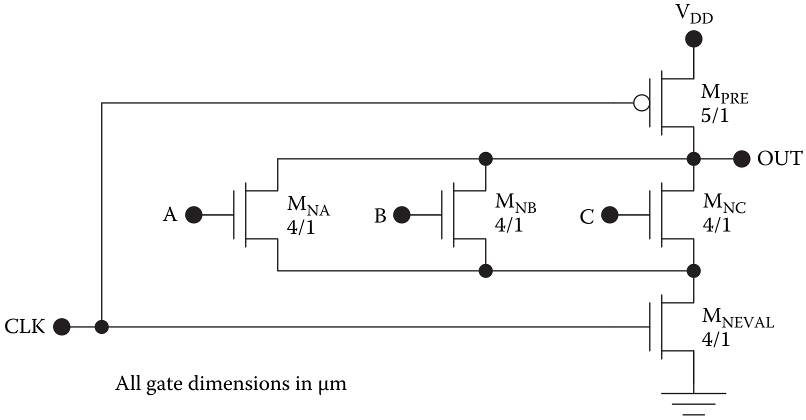


FIGURE 8.10
Dynamic CMOS NOR3 circuit.

the bottom-most transistor having the greatest width. In a dynamic CMOS NAND gate with fan-in M, the topmost transistor is connected in a series stack of M + 1 transistors. Increasing the width of this device will have a small effect on the propagation delay, which will be dominated by the remaining transistors in the stack. On the other hand, increasing the width of this device will directly scale the input capacitance for this input. The balance of these competing requirements shifts as we move toward the bottom of the stack, and, for the lower transistors, the need for higher current capability outweighs the need for low input capacitance. Therefore, based on this qualitative

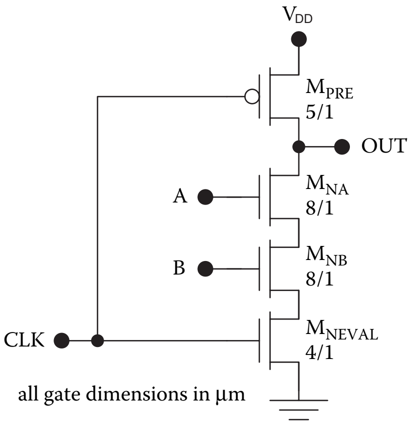


FIGURE 8.11
Dynamic CMOS NAND2 circuit.

argument, we might taper the widths of the transistors linearly from the top (least width) to the bottom of the stack (greatest width).

8.7 Alternative Form Using a p-MOS Pull-Up Network

An alternative form of dynamic CMOS using a dual p-MOS pull-up network can be realized as shown in Figure 8.12. For either the pull-down or pull-up realization of dynamic CMOS, the total number of transistors is M + 2 for a fan-in of M. Compared with static CMOS, either the pull-up or pull-down network may be eliminated with a total reduction of M – 2 transistors. All other things being equal, the pull-down realization is preferred because n-MOS transistors have greater current drive per unit gate width.

In the alternative form with a p-MOS pull-up network shown in Figure 8.12, the output is precharged to zero by an n-MOS device, whereas the evaluation is by a p-MOS transistor. Therefore, an inverted clock is used. (Precharging occurs when $\overline{CLK} = 1$ but evaluation occurs when $\overline{CLK} = 0$.) Figure 8.13 illustrates a two-way dynamic CMOS NAND circuit constructed in the alternative form using a p-MOS pull-up network to realize the logic function. The p-MOS pull-up network is the dual of the network used for the n-MOS pull-down realization. Thus, whereas a NAND2 gate would require a pull-down network with two series-connected n-MOSFETs, the dual pull-up network has two parallel-connected p-MOSFETS. In the NAND2 gate, input signals A and B are applied to p-MOS transistors connected in parallel. When $\overline{CLK} = 1$, the output is "precharged" (discharged) to zero. Then, during the

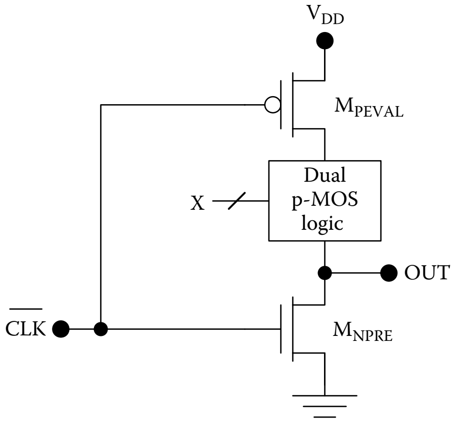


FIGURE 8.12
Alternative form of dynamic CMOS using a dual p-MOS logic network.

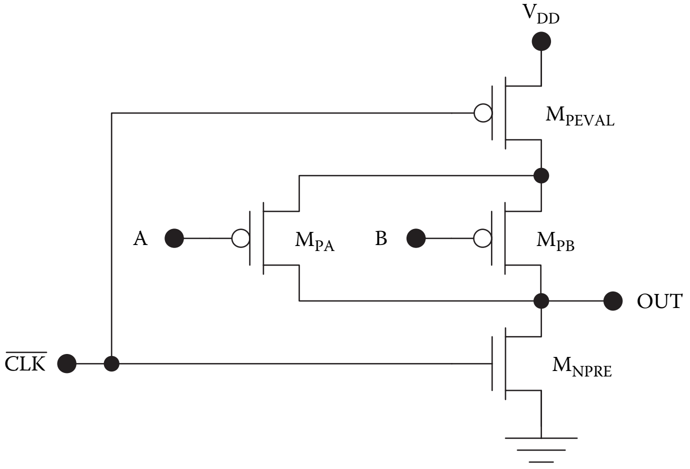


FIGURE 8.13
Dynamic CMOS NAND2 circuit using a dual p-MOS logic network.

evaluation phase, $\overline{CLK} = 0$ and the output will go high if either M_{PA} or M_{PB} turns on, that is, if either input is zero.

8.8 Cascading of Dynamic Logic Circuits

A serious limitation of the dynamic CMOS circuits described in the preceding sections is that they cannot be cascaded. For the sake of illustration, consider two cascaded dynamic CMOS inverters as shown in Figure 8.14. Both stages use the pull-down design and the circuits share a single clock signal.

Consider the behavior of this circuit with the input voltage waveform shown in Figure 8.15. At the beginning of the first precharge phase P1, the input goes low. Both nodes B and OUT precharge to V_{DD} during this precharge phase P1. During the first evaluation phase E1, node B remains at V_{DD} because A = 0. This in turn causes OUT to go to zero, and correct operation is achieved here because the voltage at node B was steady while the second stage was evaluating. However, very different results are obtained if B is not steady during the evaluation of the second stage. During the second precharge phase P2, both the B and OUT nodes precharge to V_{DD} and the input also makes a low-to-high transition. Once the second evaluation phase E2 starts, the output B begins to drop toward zero based on the logic one input to the first stage. However, the second stage begins to evaluate while the voltage is slewing at node B. Therefore, the voltage at OUT will drop considerably before the input to the second stage reaches its final value of zero, and there will be a spurious voltage output somewhere between zero and V_{DD}.

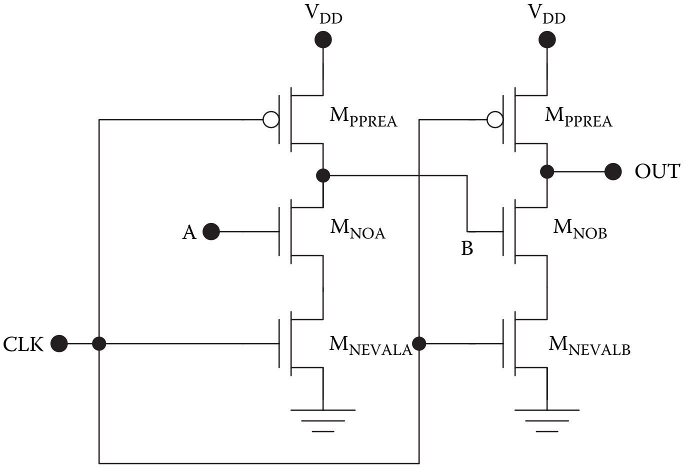


FIGURE 8.14
Cascaded dynamic CMOS inverter circuits.

This problem only comes about when the input voltage to a pull-down dynamic CMOS circuit is making a high-to-low transition during the evaluation phase. However, because all output nodes are precharged to V_{DD}, they will all either stay at V_{DD} or make a high-to-low transition. Therefore, if we insert a static inverter after each dynamic logic circuit, the inverted outputs will never make high-to-low transitions during an evaluation phase, and the problem described above may be avoided entirely.

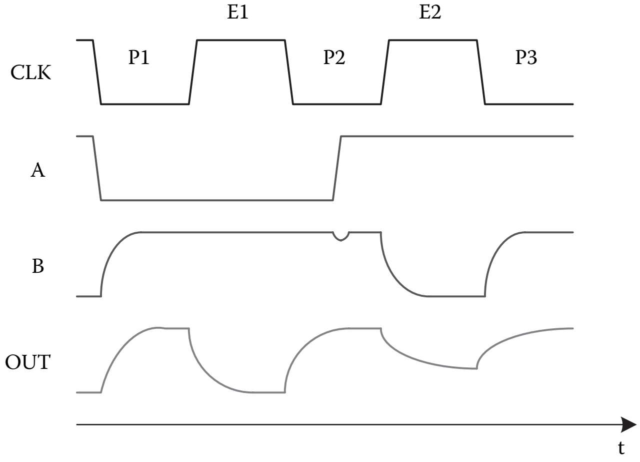


FIGURE 8.15
Example timing diagram for the cascaded dynamic CMOS inverters of Figure 8.14.

8.9 Domino Logic

In the previous section, we showed that pull-down dynamic CMOS logic circuits cannot be cascaded using a single clock signal. This is because all of the stages attempt to evaluate simultaneously, when the global clock signal goes high. During the evaluation phase, one or more of the output nodes of the dynamic circuits may make a high-to-low transition. If this signal is then fed to another dynamic CMOS circuit that is evaluating, this other gate will produce a spurious output voltage.

The problem described above may be avoided if a static CMOS inverter is placed after each dynamic CMOS gate. Then the inverted outputs will either stay constant or make low-to-high transitions during an evaluation phase. To use these inverted output signals, we can replace the logic blocks with their duals (replace parallel-connected transistors with series-connected devices and vice versa). These design changes result in a type of circuitry known as domino logic [5], which retains the speed and density advantages of dynamic CMOS, allows unlimited cascading of logic stages, and still uses a simple clocking scheme. The propagation of signals from one stage to the next is similar to the action of falling dominos and gives rise to the name of this circuit family. It takes N clock cycles for signals to propagate through N cascaded stages of domino logic.

Figure 8.16 illustrates a generalized domino logic gate in which $\overline{X}$ is the input vector. With a fan-in of M inputs, the dual n-MOS logic block requires M n-MOSFETs so the total number of transistors is M + 4, including the two transistors in the static CMOS inverter. Therefore, a domino logic circuit

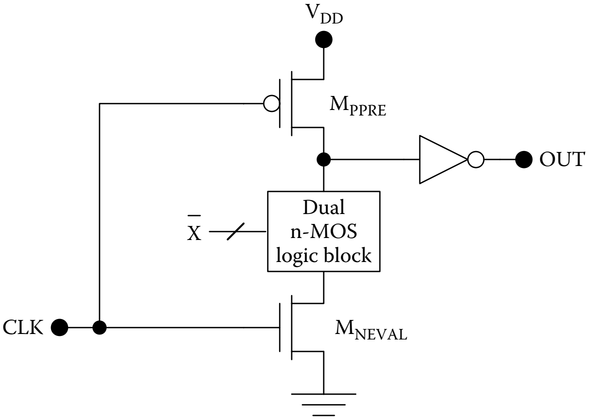


FIGURE 8.16
General domino logic circuit.

requires fewer transistors than a static CMOS gate if there are more than four inputs.

In the two-input OR gate of Figure 8.17, the dual n-MOS logic network uses series-connected n-MOSFETs and inverted inputs. This realization stems from use of DeMorgan's theorem: $A+B=\overline{\overline{A+B}}=\overline{\overline{A}\cdot\overline{B}}$. Figure 8.18 depicts another domino logic gate that realizes the function $Y=A+BC$.

One disadvantage of standard domino logic circuits is that they are all non-inverting. If inverting functions are needed, they can be implemented using static CMOS gates but with some loss of the speed and packing density advantages provided by domino logic.

8.10 Multiple-Output Domino Logic

In domino logic circuits (or other dynamic gate circuits), it is possible to avoid problems of charge sharing by precharging all intermediate soft nodes to V_{DD} using multiple precharge transistors. Doing so also makes it possible to use these intermediate soft nodes to create additional logic functions. For example, Figure 8.19 shows a two-way domino logic OR gate in which the intermediate soft node at the drain of M_{NB} is precharged by M_{PPRE2} and forms the basis for the output Y2. In the example shown, the circuit implements the

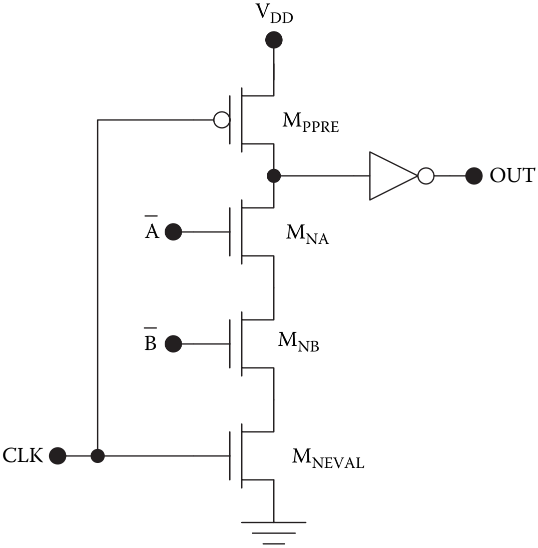


FIGURE 8.17
Domino two-way OR circuit.

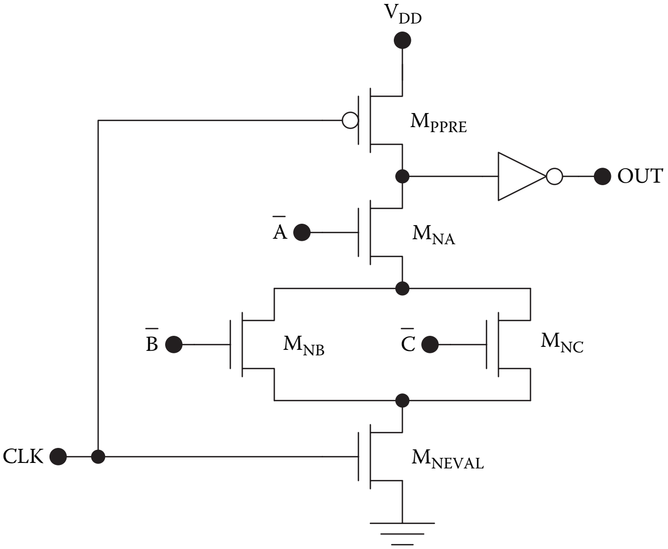


FIGURE 8.18
Domino circuit to perform the logic function $Y = A + BC$.

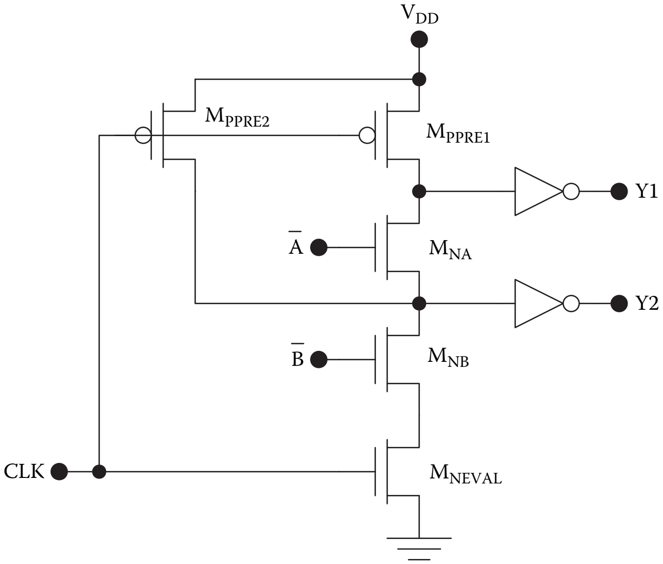


FIGURE 8.19
Domino logic circuit with two outputs. $Y1 = A + B$ and $Y2 = \overline{B}$.

two functions $Y1 = A + B$ and $Y2 = \overline{B}$. Although each output node requires a precharge transistor, only a single evaluation transistor is needed.

Another example of a domino logic circuit with multiple outputs is the Manchester carry chain (MCC) shown in Figure 8.20 [6]. The functions implemented by the MCC circuit are

$$C1 = G1 + P1 \cdot C0 \,, \tag{8.11}$$

$$C2 = G2 + P2 \cdot G1 + P2 \cdot P1 \cdot C0 \,, \tag{8.12}$$

$$C3 = G3 + P3 \cdot G2 + P3 \cdot P2 \cdot G1 + P3 \cdot P2 \cdot P1 \cdot C0 \,, \tag{8.13}$$

and

$$C3 = G4 + P4 \cdot G3 + P4 \cdot P3 \cdot G2 + P4 \cdot P3 \cdot P2 \cdot G1 + P3 \cdot P2 \cdot P1 \cdot C0 \,. \tag{8.14}$$

This practical circuit generates the carry bits C1, C2, C3, and C4 in a four-stage carry-look-ahead adder, where $G_i = A_i \cdot B_i$ and $P_i = A_i \oplus B_i$ and A_i and B_i are the bits of the words being added.

8.11 Zipper Logic

In dynamic CMOS circuits using n-MOS pull-down networks and a single clock, logic circuits cannot be cascaded because a high-to-low transition at the output of one gate will produce a spurious voltage at the output of a fan-out gate that evaluates at the same time. On the other hand, dynamic CMOS circuits using dual p-MOS logic networks will only produce spurious voltage outputs if their inputs make low-to-high transitions during the evaluation phase. Hence, during an evaluation phase, we should avoid a high-to-low input transition to a dynamic CMOS gate with an n-MOS logic network but a low-to-high input transition to a dynamic CMOS gate with a p-MOS logic network. Whereas domino logic uses all n-MOS logic stages with inverters to solve this problem, another solution is to alternate n-MOS logic stages with p-MOS logic stages. This approach, called *zipper logic* [7], is successful because the output of a n-MOS stage will either stay the same or make a high-to-low transition, but the output of a p-MOS stage will either stay low or make a low-to-high transition during evaluation. Figure 8.21 shows four cascaded stages of zipper logic to illustrate the design principle.

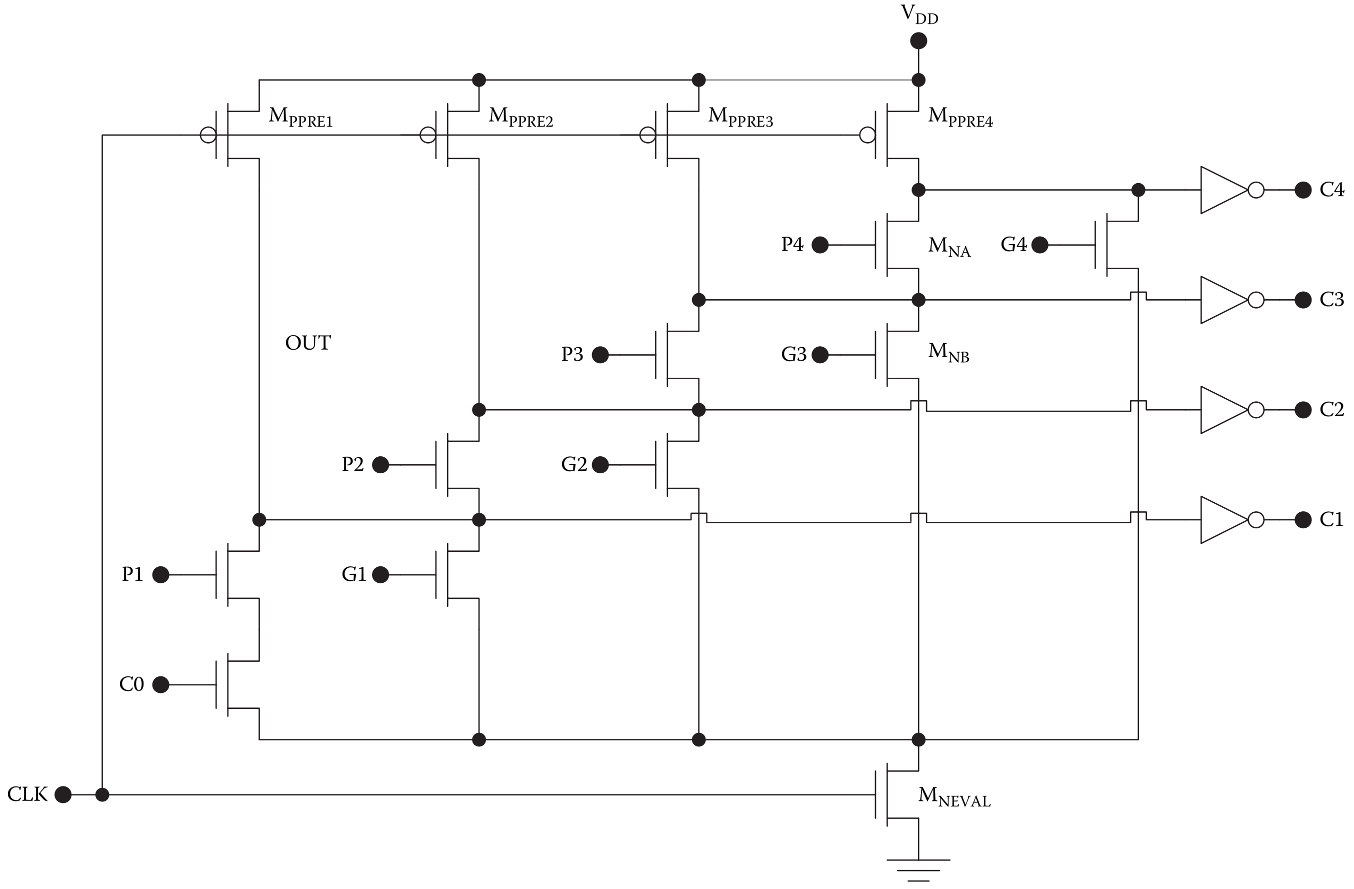


FIGURE 8.20
Domino logic MCC.

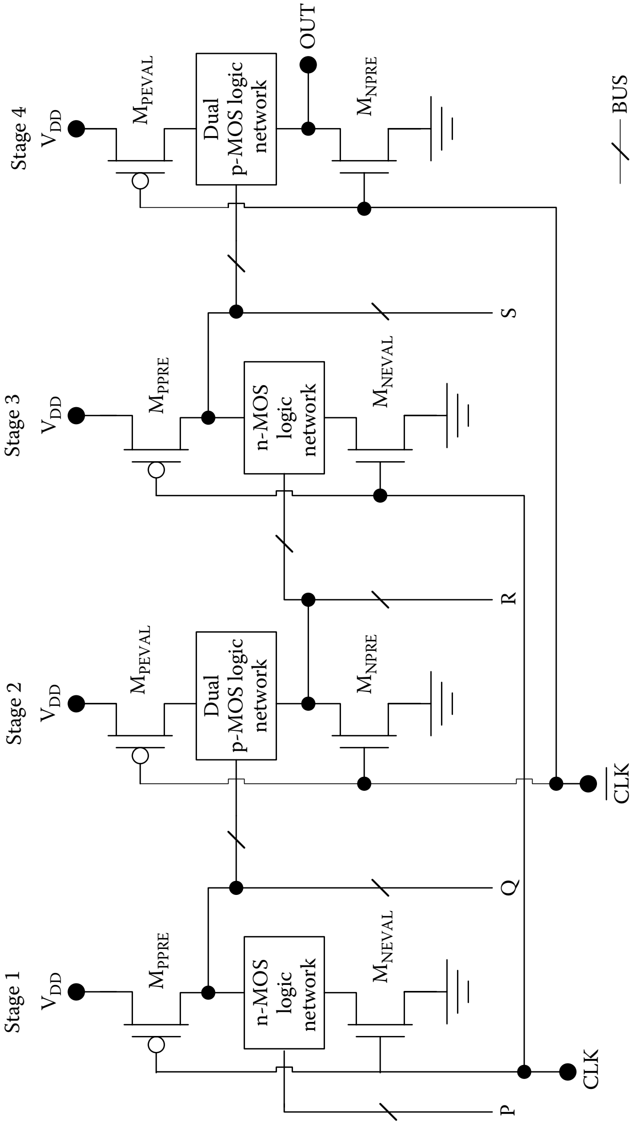


FIGURE 8.21
Four stages of cascaded zipper logic.

8.12 Dynamic Pass Transistor Circuits

In Chapter 5, we considered the use of pass transistors to create static pass transistor logic circuits. We can also use pass transistors in dynamic circuits to control the transmission of signals from one stage of logic to another. As an example, consider the three-stage shift register shown in Figure 8.22. This circuit uses a *two-phase clocking scheme* in which the two clock pulses φ_1 and φ_2 are nonoverlapping. Each stage comprises a static inverter circuit with an n-MOS pass transistor connected to its input. When φ_1 goes high, the first pass transistor M_1 turns on, transferring the input voltage to inverter 1. The output of inverter 1 settles based on this input voltage, which is also stored in the input capacitance C_{in1}. When φ_2 goes high, the voltage on C_{out1} is transferred to the input of gate 2 (which evaluates accordingly) and also stored in its input capacitance. Next the phase φ_1 goes high again, so that the output

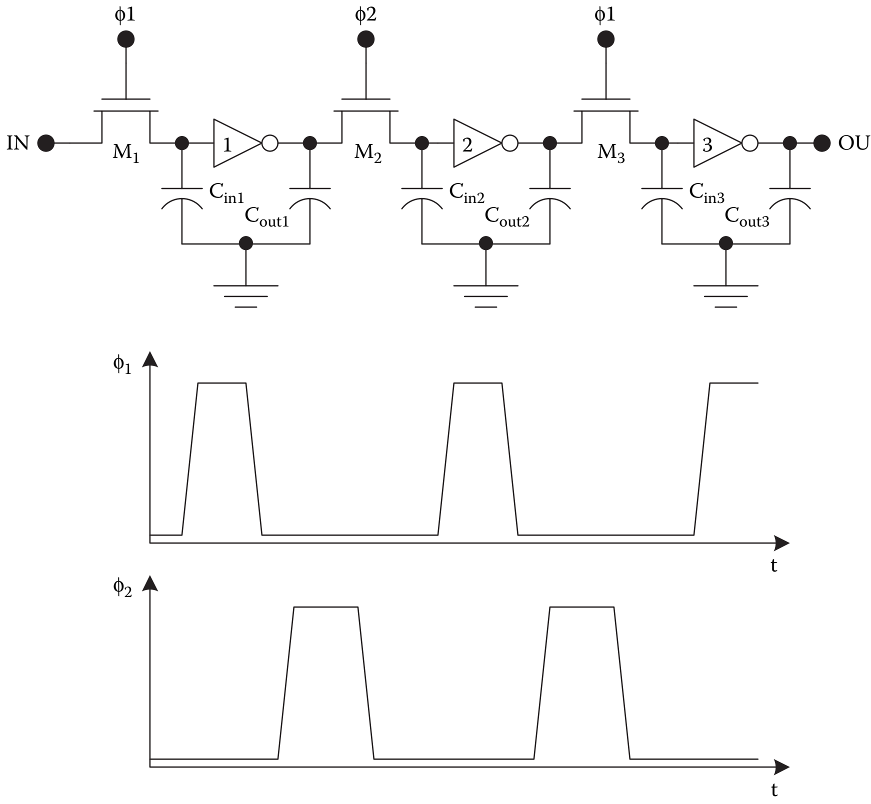


FIGURE 8.22
Dynamic one-bit shift register with three stages.

voltage from inverter 2 is transferred to the input of inverter 3. At the same time, another bit may be pipelined into the shift register at its input.

The previous example represents the simplest type of dynamic pass transistor circuit, in which only a single bit is passed from one stage to the next. In general, an arbitrary number of digital signals may be transferred from one stage to the next, but each must have its own pass transistor. Figure 8.23 shows an example of a three-stage dynamic pass transistor circuit involving complex logic functions of five input signals. When φ_1 goes high, inputs A, B, and C are transferred to the stage 1 logic. When φ_2 goes high, the outputs of stage 1 as well as inputs D and E are transferred to the stage 2 logic. During the next φ_1 pulse, the output of stage 2 and input F are transferred to the stage 3 logic. The node capacitances, although omitted from the figure for simplicity, are necessary for the correct operation of the circuit. The two clock pulses are made non-overlapping so that a particular stage will never be settling while being read by the next stage. This ensures an orderly transfer of data from one stage to the next in this *sequential circuit*.

8.12.1 Logic "1" Transfer Delay t_1

When an n-MOS pass transistor is used to transfer logic 1 to the soft node of a dynamic circuit, it will take a finite time (the charge-up time t_1) for the soft node voltage to settle, and this places a lower limit on the clock pulse width. Consider the simplified circuit of Figure 8.24, in which logic "1" is transferred from the input node to a soft node with an equivalent lumped capacitance of C_X. To simplify the analysis, it will be assumed that the clock φ makes an abrupt low-to-high transition at t = 0 and that the body effect may be neglected for the pass transistor M_1. We will further assume that the initial voltage on the soft node is zero.

During the charge-up process, the pass transistor will operate in saturation because $V_{GS} = V_{DS}$ with V_{DD} applied at its gate. Therefore,

$$C_X \frac{dV_X}{dt} = \frac{K_N}{2}\left(V_{DD} - V_X - V_{TN}\right)^2 \tag{8.15}$$

where K_N and V_{TN} are the device transconductance parameter and threshold voltage for the pass transistor, respectively. Rearranging and integrating both sides, we obtain

$$\int_0^t d\tau = \frac{2C_X}{K_N}\int_0^{V_X} \frac{dV}{\left(V_{DD} - V - V_{TN}\right)^2} \tag{8.16}$$

so that

$$t = \frac{2C_X}{K_N}\left[\left(\frac{1}{V_{DD} - V_X - V_{TN}}\right) - \left(\frac{1}{V_{DD} - V_{TN}}\right)\right]. \tag{8.17}$$

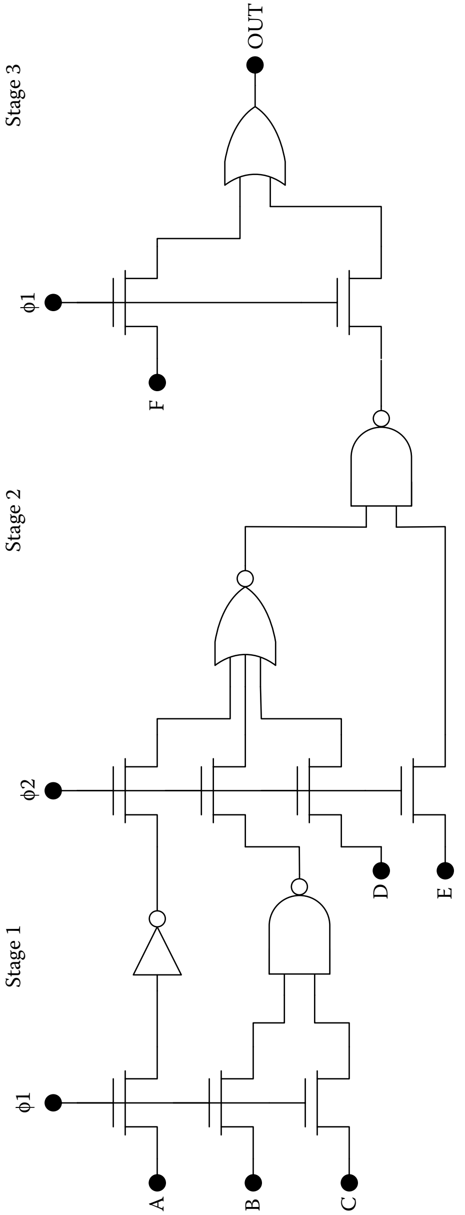


FIGURE 8.23
A three-stage dynamic pass transistor circuit.

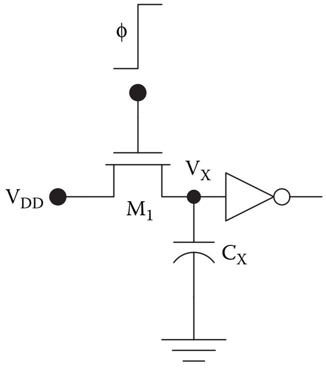


FIGURE 8.24
Logic "1" transfer in a pass transistor circuit.

The solution for V_X is

$$V_X(t) = \frac{V_{DD} - V_{TN}}{1 + \left(\dfrac{2C_X}{K_N\left(V_{DD} - V_{TN}\right)t}\right)}. \tag{8.18}$$

If we define the charge-up time (the time for logic "1" transfer t1) to be the time it takes for V_X to reach 90% of its final value, $V_X\left(t_1\right) \equiv 0.9\left(V_{DD} - V_{TN}\right)$, then

$$t_1 \approx \frac{18C_X}{K_N\left(V_{DD} - V_{TN}\right)}. \tag{8.19}$$

In this analysis, we neglected the body effect in the pass transistor. However, the source of M_1 is not grounded but at a positive voltage V_X; as a consequence, the threshold voltage will be reduced and Equation 8.19 therefore overestimates t_1 somewhat.

8.12.2 Logic "0" Transfer Delay t_0

When an n-MOS pass transistor is turned on to transfer logic "0" to the soft node of a dynamic circuit, there is a charge-down delay time t_0. To analyze this delay, we will use the simplified circuit of Figure 8.25 with the assumptions that the clock φ makes an abrupt low-to-high transition at t = 0 and the soft node voltage is initially V_{DD}. There is no body effect in this case because the source of the pass transistor is grounded.*

* For the symmetric n-MOS pass transistor, the source and drain regions can exchange roles based on the biasing and the direction of current flow. For the logic "1" transfer, the conventional current flow was left to right, so the right-hand contact played the role of source. For logic "0" transfer, the conventional current flow is from right to left, so the left-hand contact region acts as the source.

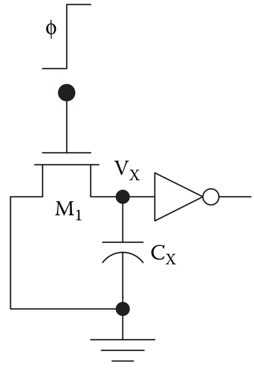


FIGURE 8.25
Logic "0" transfer in a pass transistor circuit.

The pass transistor will operate in saturation until V_X drops to $V_{DD} - V_{TN}$ at $t = t_{01}$. During this time interval,

$$C_X \frac{dV_X}{dt} = \frac{K_N}{2}\left(V_{DD} - V_{TN}\right)^2 \tag{8.20}$$

so the time delay associated with saturated operation of the pass transistor is

$$t_{01} = \frac{C_X}{K_N}\frac{2V_{TN}}{\left(V_{DD} - V_{TN}\right)^2}\ . \tag{8.21}$$

After the soft node voltage drops below $V_{DD} - V_{TN}$, the pass transistor will be linear. We will define the logic "0" transfer delay to be the time required for the soft node voltage to reach 10% of its limiting value. Then the time delay for linear operation of the pass transistor is

$$\begin{aligned} t_{02} &= -C_X \int_{V_{DD}-V_{TN}}^{V_{DD}/10} \frac{dV_X}{K_N\left[\left(V_{DD} - V_{TN}\right)V_X - V_X^2/2\right]} \\ &= -\frac{C_X}{K_N\left(V_{DD} - V_{TN}\right)}\left[\ln\left(\frac{V_X}{V_X - 2\left(V_{DD} - V_{TN}\right)}\right)\right|_{V_{DD}-V_{TN}}^{V_{DD}/10} \\ &= \frac{C_X}{K_N\left(V_{DD} - V_{TN}\right)}\ln\left(\frac{19V_{DD} - 20V_{TN}}{V_{DD}}\right). \end{aligned} \tag{8.22}$$

The logic "0" transfer time is given by the sum $t_0 = t_{01} + t_{02}$ so

$$t_0 = \frac{C_L}{K_N\left(V_{DD} - V_{TN}\right)}\left[\frac{2V_{TN}}{\left(V_{DD} - V_{TN}\right)} + \ln\left(\frac{19V_{DD} - 20V_{TN}}{V_{DD}}\right)\right]. \tag{8.23}$$

Example 8.3 Pass Transistor Transfer Delays t_0 and t_1

Estimate the logic "0" and logic "1" transfer times for a pass transistor connected to a CMOS inverter as shown in the circuit diagram of Figure 8.26. The physical layout of the circuit is shown in Figure 8.27. Assume that $V_{DD} = 2.5V$, $V_{TN} = |V_{TP}| = 0.5V$, t_{OX} = 9 nm, x_j = 0.1 µm, and L_{OV} = 0.1 µm. Assume abrupt source and drain junctions with $N_d = 10^{19}$ cm^{-3}. The substrate doping concentration is $N_a = 10^{16}$ cm^{-3}, whereas the sidewall (channel stopper) doping is $N_a =$ 10^{17} cm^{-3}.

Solution: For the pass transistor,

$$k_N' = \frac{\mu_n \varepsilon_{OX}}{t_{OX}} = \frac{\left(580cm^2/Vs\right)(3.9)\left(8.85\times10^{-14}F/cm\right)}{9\times10^{-7}cm} = 220\mu A/V^2$$

and

$$K_N = k_N'\left(\frac{W_N}{L_N}\right) = 220\mu A/V^2\left(\frac{4}{2}\right) = 440\mu A/V^2.$$

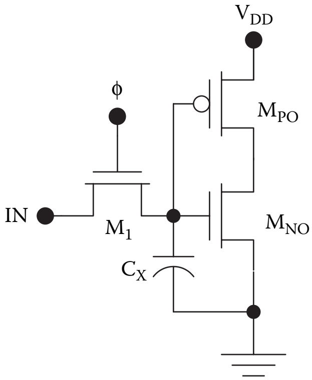


FIGURE 8.26
Pass transistor driving a CMOS inverter. C_X represents the soft node capacitance that is associated primarily with the oxide capacitances of M_{NO} and M_{PO}.

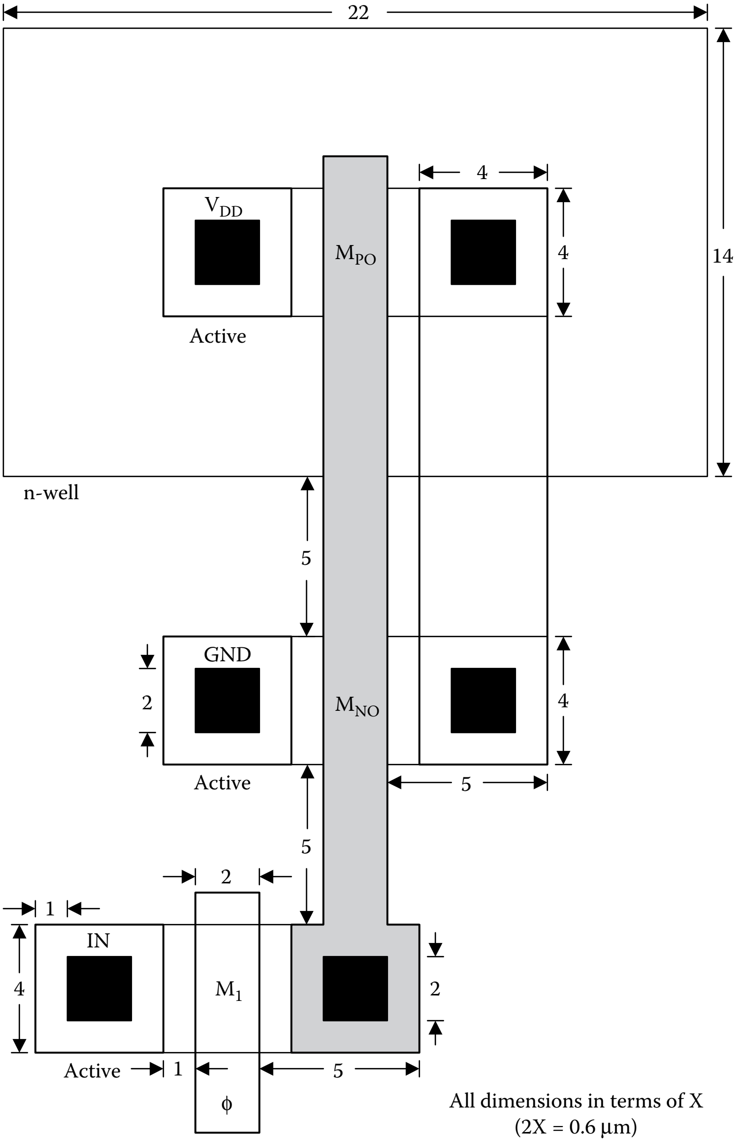


FIGURE 8.27
Physical layout for a pass transistor M_1 driving a CMOS inverter comprising the transistors M_{NO} and M_{PO}. All dimensions are given in terms of X, where $2X = 0.6\mu m$.

The oxide capacitance per unit area is

$$C_{ox} = \varepsilon_{ox} / t_{ox} = 3.83 \times 10^{-15} F / \mu m^2$$

and the input capacitance to the CMOS inverter is

$$\begin{aligned} C_{in} &= C_{gN} + C_{gP} = C_{ox}\left(W_N L_N + 2W_N L_{OV}\right) + C_{ox}\left(W_P L_P + 2W_P L_{OV}\right) \\ &= 3.83\times10^{-15} F/\mu m^2\left[(1.2)(0.6)+2(1.2)(0.1)\right] \\ &\quad + 3.83\times10^{-15} F/\mu m^2\left[(1.2)(0.6)+2(1.2)(0.1)\right] \\ &= 3.7fF + 3.7fF = 7.4fF. \end{aligned}$$

For the p-n junction of the pass transistor (the S/D junction connected to the soft node, on the right hand side of the device) we should consider the bottom and sidewall capacitances. Considering worst-case zero-bias values, for the bottom junction we have

$$V_{bi} = \frac{kT}{q}\ln\left(\frac{N_a N_d}{n_i^2}\right) = (0.026V)\ln\left(\frac{\left(10^{16}cm^{-3}\right)\left(10^{19}cm^{-3}\right)}{\left(1.45\times10^{10}cm^{-3}\right)^2}\right) = 0.88V\,,$$

$$W = \sqrt{\frac{2\varepsilon_s V_{bi}}{qN_a}} = \sqrt{\frac{2(11.9)\left(8.85\times10^{-14}F/cm\right)(0.88V)}{\left(1.602\times10^{-19}C\right)\left(10^{16}cm^{-3}\right)}} = 0.34\times10^{-4}cm\,, \\ = 0.34\mu m$$

and the zero-bias depletion capacitance per unit area is

$$C_{bot0} = \frac{(11.9)\left(8.85\times10^{-14}F/cm\right)}{0.34\times10^{-4}cm} = 0.31\times10^{-15}F/\mu m^2\,.$$

For the p-n junctions at the sidewalls,

$$V_{bi} = \frac{kT}{q}\ln\left(\frac{N_a N_d}{n_i^2}\right) = (0.026V)\ln\left(\frac{\left(10^{17}cm^{-3}\right)\left(10^{19}cm^{-3}\right)}{\left(1.45\times10^{10}cm^{-3}\right)^2}\right) = 0.94V\,,$$

$$W = \sqrt{\frac{2\varepsilon_s V_{bi}}{qN_a}} = \sqrt{\frac{2(11.9)\left(8.85\times10^{-14}F/cm\right)(0.94V)}{\left(1.602\times10^{-19}C\right)\left(10^{17}cm^{-3}\right)}} = 0.111\times10^{-4}cm\,, \\ = 0.111\mu m$$

and with zero bias,

$$C_{sw0} = \frac{(11.9)\left(8.85\times10^{-14}F/cm\right)}{0.111\times10^{-4}cm} = 0.95\times10^{-15}F/\mu m^2\,.$$

The total junction capacitance of the pass transistor connected to the soft node is

$$\begin{aligned} C_D &\approx WL_D C_{bot} + x_j\left(2L_D + W\right)C_{sw} \\ &= \left(1.2\mu m\right)\left(5\mu m\right)\left(0.31fF/\mu m^2\right) \\ &\quad +\left(0.1\mu m\right)\left(3.0\mu m + 1.2\mu m\right)\left(0.95fF/\mu m^2\right) \\ &= 1.9fF + 0.4fF = 2.3fF. \end{aligned}$$

The soft node capacitance is the sum of the junction capacitance of the pass transistor and the input capacitance for the CMOS inverter:

$$C_X = 7.4fF + 2.3fF = 9.7fF \cdot$$

The logic "0" transfer delay time is

$$\begin{aligned} t_0 &= \frac{C_L}{K_N\left(V_{DD} - V_{TN}\right)}\left[\frac{2V_{TN}}{\left(V_{DD} - V_{TN}\right)} + \ln\left(\frac{19V_{DD} - 20V_{TN}}{V_{DD}}\right)\right] \\ &= \frac{9.7\times10^{-15}F}{440\times10^{-6}A/V^2\left(2.5V - 0.5V\right)}\left[\frac{2\left(0.5V\right)}{\left(2.5V - 0.5V\right)} + \ln\left(\frac{19\left(2.5V\right) - 20\left(0.5V\right)}{2.5V}\right)\right] \\ &= 35ps. \end{aligned}$$

and the logic "1" transfer delay time is

$$t_1 \approx \frac{18C_X}{K_N\left(V_{DD} - V_{TN}\right)} = \frac{18\left(9.7\times10^{-15}F\right)}{440\times10^{-6}A/V^2\left(2.5V - 0.5V\right)} = 200ps \cdot$$

The logic "1" transfer delay time is substantially longer than the logic "0" transfer delay time.

8.13 CMOS Transmission Gate Circuits

In the dynamic pass transistor circuits described in the previous section, the n-MOS pass transistors act essentially as transmission gates. However, the maximum voltage that can be transferred by an n-MOS pass transistor is $V_{DD} - V_{TN}$; that is, a logic one voltage is degraded by the threshold voltage of the pass transistor. This problem may be avoided by replacing the n-MOS pass transistors with CMOS transmission gates, as shown in Figure 8.28 for the case of a three-stage CMOS transmission gate shift register. If a true two-phase clocking scheme is adopted, it is necessary to distribute four separate clock signals. This is because the transmission gates require complementary clock signals. The first transmission gate T1 is driven by ϕ_1 and $\overline{\phi_1}$. However, ϕ_1 and $\overline{\phi_1}$ do not constitute a non-overlapping two-phase clock scheme because (1) these signals have finite rise and fall times and (2) $\overline{\phi_1}$ is produced by inverting ϕ_1 so it exhibits a time delay introduced by the inverting circuit.

As with the pass transistor circuits, CMOS transmission gate logic may be extended to more complex logic functions. The logic blocks are implemented as static logic circuits, implemented in CMOS or any other type of static logic circuitry.

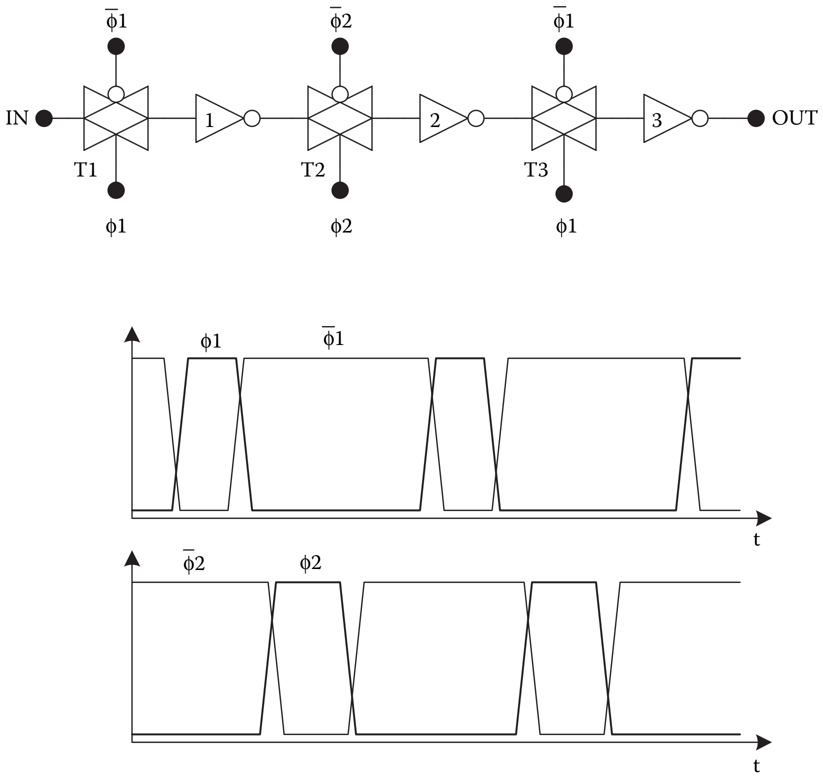


FIGURE 8.28
Dynamic CMOS transmission gate shift register.

8.14 SPICE Demonstrations

For the purpose of illustration, simulations were performed using Cadence Capture CIS 10.1.0 PSpice (Cadence Design Systems). The level 1 MOS transistor model parameters given in Tables 8.1 and 8.2 were used unless otherwise noted. The process transconductance parameters were calculated assuming an oxide thickness of 9 nm. For n-MOSFETS,

$$KP = \frac{(3.9)\left(8.85\times10^{-14}F/cm\right)\left(580cm^2V^{-1}s^{-1}\right)}{9\times10^{-7}cm} = 222\mu A/V^2, \tag{8.24}$$

and for p-MOSFETS,

$$KP = \frac{(3.9)\left(8.85\times10^{-14}F/cm\right)\left(230cm^2V^{-1}s^{-1}\right)}{9\times10^{-7}cm} = 88\mu A/V^2, \tag{8.25}$$

The overlap capacitances per unit gate width were determined with the assumption that $L_{OV} = 0.1\mu m$:

$$\begin{aligned} CGSO &= \frac{(3.9)\left(8.85\times10^{-14}F/cm\right)\left(0.1\times10^{-4}cm\right)}{9\times10^{-7}cm} \\ &= 3.8pF/cm = 0.38nF/m \end{aligned} \tag{8.26}$$

and

$$\begin{aligned} CGDO &= \frac{(3.9)\left(8.85\times10^{-14}F/cm\right)\left(0.1\times10^{-4}cm\right)}{9\times10^{-7}cm} \\ &= 3.8pF/cm = 0.38nF/m. \end{aligned} \tag{8.27}$$

The body effect coefficient was calculated from

$$\begin{aligned} GAMMA &= \frac{\sqrt{2q\varepsilon_{Si}N_a}}{C_{ox}} \\ &= \frac{\sqrt{2\left(1.602\times10^{-19}C\right)(11.9)\left(8.85\times10^{-14}F/cm\right)\left(10^{16}cm^{-3}\right)}}{(3.9)\left(8.85\times10^{-14}F/cm\right)/9\times10^{-7}cm} \\ &\approx 0.15V^{1/2}. \end{aligned} \tag{8.28}$$

TABLE 8.1

n-MOS Level 1 SPICE Parameters

Parameter	Value	Units
KP	222u	A/V^2
VTO	0.5	V
GAMMA	0.15	V$^{1/2}$
PHI	0.7	V
LAMBDA	0.05	
TOX	9n	m
NSUB	1E16	cm^3
UO	580	cm^2/Vs
CGSO	0.38n	F/m
CGDO	0.38n	F/m

SPICE Example 8.1 Dynamic CMOS Inverter

A transient simulation was performed for the dynamic CMOS inverter of Figure 8.29 to show its basic operation. The clock pulse source was set up with V1 = 2.5 V, V2 = 0, TD = 0, TR = 0.1 ns, TF = 0.1 ns, PW = 0.8 ns, and PER = 2 ns. The input voltage was set up with V1 = 2.5 V, V2 = 0, TD = 0, TR = 0.1 ns, TF = 0.1 ns, PW = 2.5 ns, and PER = 20 ns. The transient results are shown in Figure 8.30. On the first high-to-low transition of the clock, the output node precharges to V_{DD} and remains at this level while the input voltage is low. When the input voltage makes the low-to-high transition at t = 2.6 ns, the output voltage drops slightly attributable to charge sharing. During the next low-to-high transition of the clock at t = 3 ns, the output evaluates to 0 with a fall time of ~ 0.7 ns.

TABLE 8.2

p-MOS Level 1 SPICE Parameters

Parameter	Value	Units
KP	88u	A/V^2
VTO	– 0.5	V
GAMMA	0.15	V$^{1/2}$
PHI	0.7	V
LAMBDA	0.05	
TOX	9n	m
NSUB	1E16	cm^{-3}
UO	230	cm2/Vs
CGSO	0.38n	F/m
CGDO	0.38n	F/m

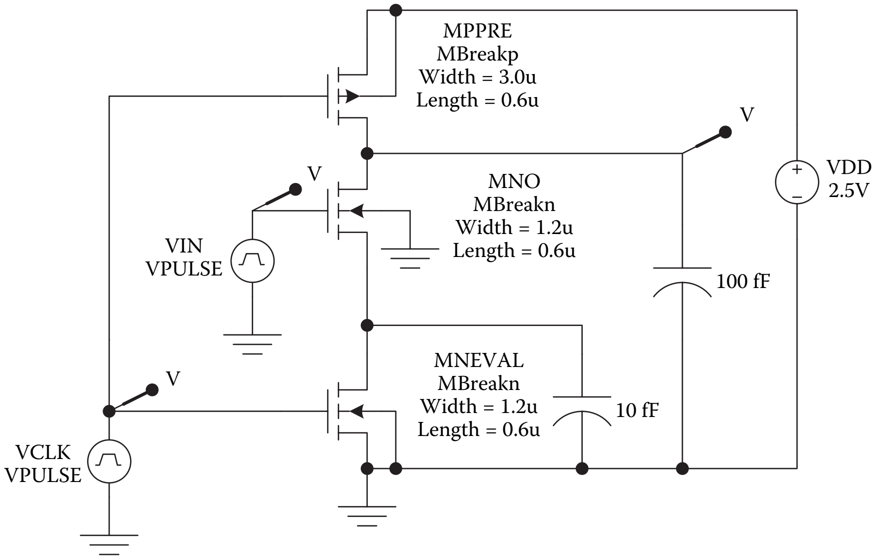


FIGURE 8.29
Dynamic CMOS inverter for determination of the transient response.

SPICE Example 8.2 Cascaded Dynamic CMOS Gates

To demonstrate the problem of cascading dynamic CMOS gates, the transient response was obtained for two cascaded dynamic CMOS inverters as shown in Figure 8.31. The clock pulse source was set up with V1 = 2.5 V, V2 = 0, TD = 0, TR = 0.1 ns, TF = 0.1 ns, PW = 0.8 ns, and PER = 2 ns. The input voltage was set up with V1 = 2.5 V, V2 = 0, TD = 0, TR = 0.1 ns, TF = 0.1 ns, PW = 2.5 ns, and PER =

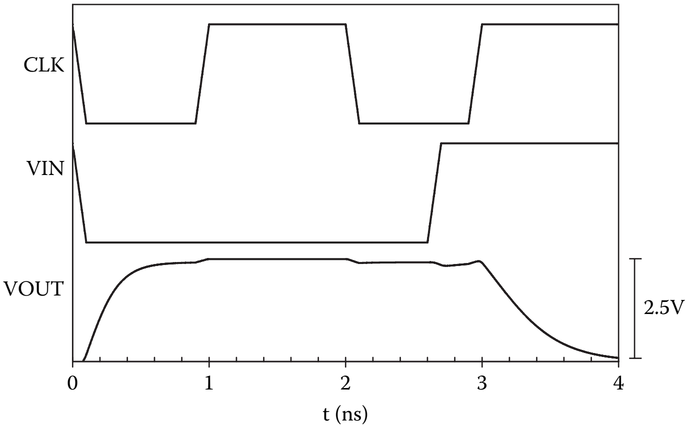


FIGURE 8.30
Transient response for the dynamic CMOS inverter of Figure 8.29.

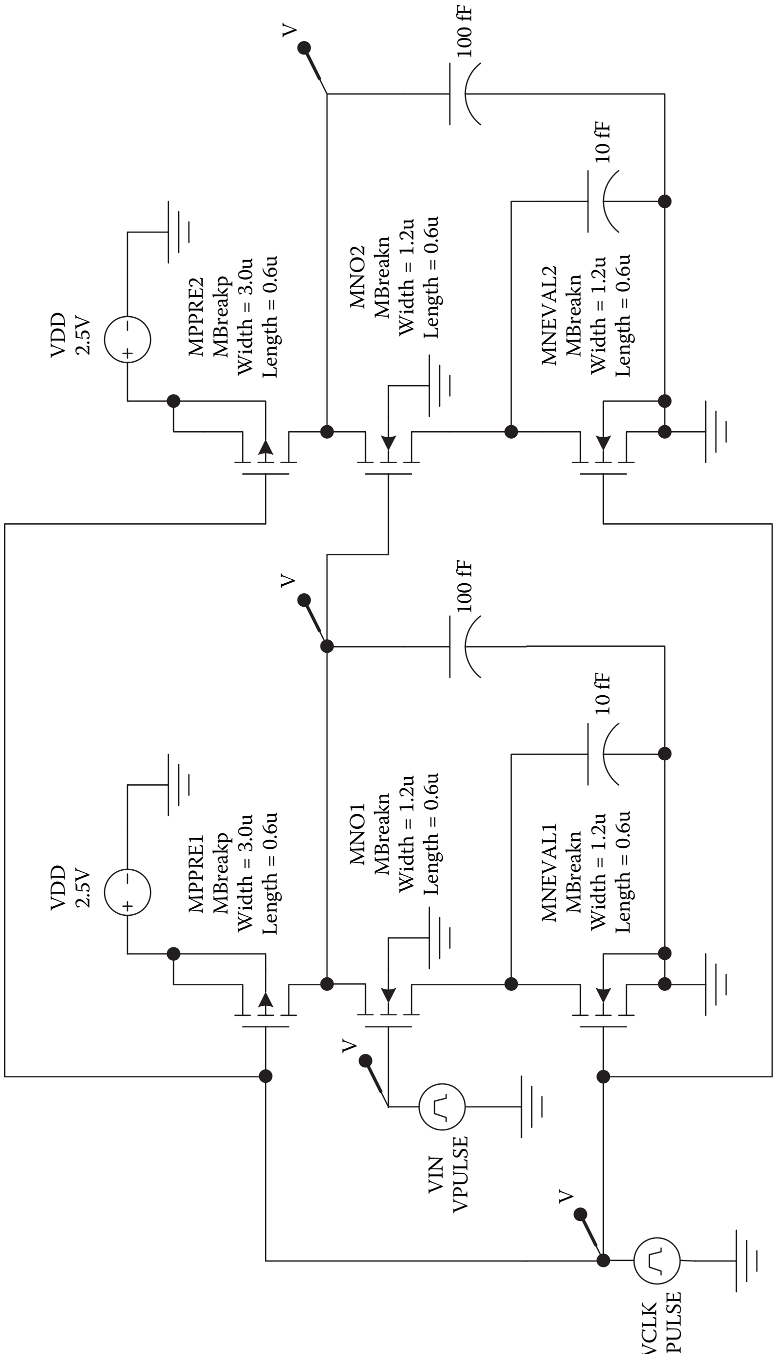


FIGURE 8.31
Cascaded dynamic CMOS inverters for the determination of the transient response.

20 ns. During the first precharge phase P1, both OUT1 and OUT2 precharge to V_{DD}. During the first evaluation phase E1, OUT2 evaluates to 0. The input voltage makes a low-to high transition during the second precharge phase P2, so that OUT1 evaluates to zero during the second evaluation phase P2. However, the second inverter is evaluating while OUT1 is changing, so it ends up with a non-valid output voltage of approximately 1.5 V. Therefore, for a dynamic logic gate with an n-MOS logic network, it must be ensured that the input to a dynamic logic gate does not make a high-to-low transition while the gate is evaluating (Figure 8.32).

8.15 Practical Perspective

For practical perspective articles, see the dynamic website at http://www.engr.uconn.edu/ece/books/ayers.

8.16 Summary

Dynamic logic circuits take advantage of the capacitances at circuit nodes to store charge; this allows the implementation of logic functions with fewer

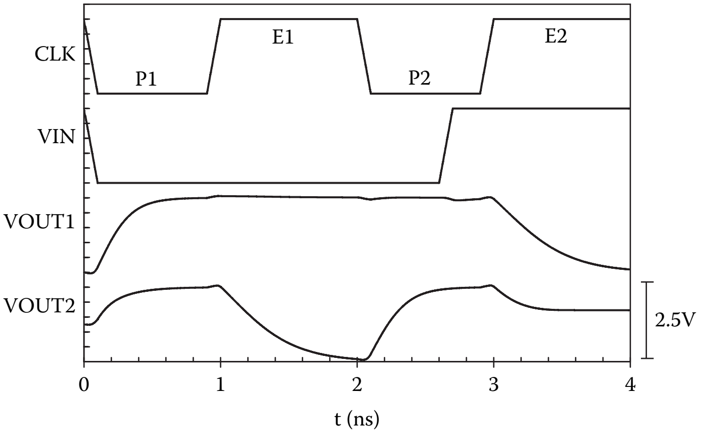


FIGURE 8.32
Transient response for the cascaded dynamic CMOS inverters of Figure 8.31.

transistors, increasing the packing density and decreasing the dissipation. In dynamic CMOS circuits, a soft output node is precharged to either zero or V_{DD} during the precharge phase of the clock and then evaluated based on the inputs to an n-MOS logic network or a dual p-MOS logic network during the evaluation phase of the clock. Dynamic gates of this type may not be cascaded directly because, if a dynamic gate tries to evaluate while its input is changing (in other words, while the previous stage is evaluating), the result may be an incorrect output voltage. Special dynamic logic styles have been designed to avoid this problem, including domino logic and zipper logic.

Dynamic circuits may also be implemented using pass transistors or transmission gates, including latches and shift registers.

8.17 Exercises

E8.1. Calculate the rise and fall times for the dynamic CMOS inverter of Figure 8.33, assuming that the clock makes abrupt high-to-low and low-to-high transitions.

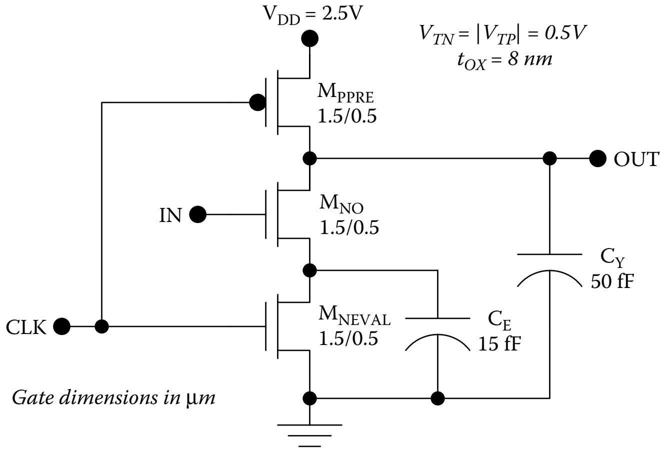


FIGURE 8.33
Dynamic CMOS inverter for determination of the rise and fall times (see Exercise E8.1).

E8.2. Estimate the charge retention time for the output node of the dynamic inverter in Figure 8.34, assuming that the subthreshold swing is 75 mV for all MOS transistors.

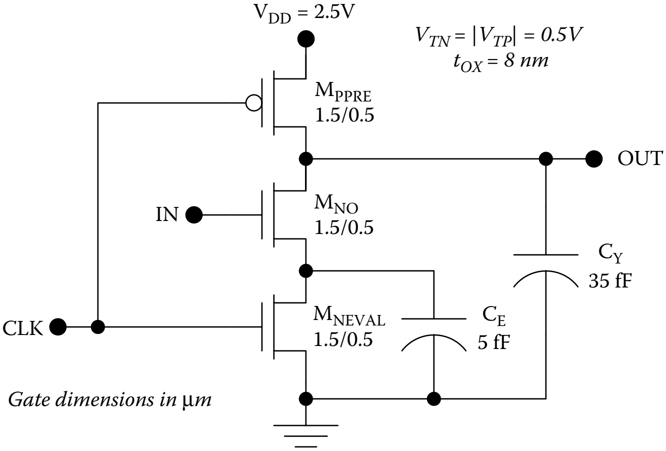


FIGURE 8.34
Dynamic circuit for the estimation of the charge retention time (see Exercise E8.2).

E8.3. Estimate the logic "1" transfer delay t_1 and the logic "0" transfer delay t_0 for the pass transistor circuit shown in Figure 8.35, assuming that the input capacitance of the CMOS inverter is the dominant component of the soft node capacitance.

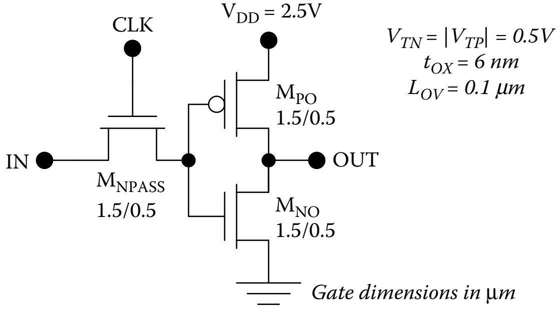


FIGURE 8.35
Pass transistor circuit for the determination of the logic "1" and logic "0" transfer times.

For additional exercise problems, see the dynamic website at http://www.engr.uconn.edu/ece/books/ayers.

References

1. Baker, R.J., *CMOS circuit design, layout, and simulation,* 2nd ed., IEEE Press, Piscataway, NJ, 2007.
2. Karnik, T., Hazucha, P., and Patel, J., Characterization of soft errors caused by single event upsets in CMOS processes. *IEEE Trans. Depend. Secure Computing*, 1, 128–143, 2004.
3. Fishburn, J.P., Clock skew optimization. *IEEE Trans. Computers*, 39, 945–951, 1990.
4. Kuo, J.B., and Chiang, C.S., Charge sharing problems in dynamic logic circuits: BiCMOS versus CMOS and a 1.5V BiCMOS dynamic logic circuit free from charge sharing problems. *IEEE Trans. Circuits Syst.*, 42, 974–977, 1995.
5. Krambeck, R.H., Lee, C.M., and Law, H.-F.S., High-speed compact circuits with CMOS. *IEEE J. Solid-State Circuits*, 17, 614–619, 1982.
6. Kernhof, J., Beunder, M.A., Hoefflinger, B., and Haas, W., High-speed CMOS adder and multiplier modules for digital signal processing in a semicustom environment. *IEEE J. Solid-State Circuits*, 24, 570–575, 1989.
7. Tong, Q., and Jha, N.K., Testing of zipper CMOS logic circuits. *IEEE J. Solid-State Circuits*, 25, 877–880, 1990.

9

Low-Power CMOS

9.1 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 by 50 times. Moreover, removal of heat from 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 tradeoff 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 just provide 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 three times the subthreshold swing [9]. However, this restriction can be lifted by the use of variable threshold CMOS (active body biasing) [10–14] or multiple threshold CMOS [15, 16]. Another solution to this problem is the use of SOI [17–44]. SOI devices have superior subthreshold swings, allowing the operation of SOI CMOS with $V_T = 0.1V$ 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. Reduction of the load capacitances can be achieved by scaling of 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 will halve the overall power dissipation. Charge recycling (reuse of electrical charges for more than one logic operation) is achieved in adiabatic logic circuits [45–47], 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, although they are beyond the scope of the present book.

9.2 Low-Voltage CMOS

As described in Chapter 6, the dissipation in a CMOS gate is given by

$$P = \underbrace{P_{subthreshold} + P_{leakage}}_{P_{DC}} + \underbrace{P_{sc} + P_{switch}}_{P_{AC}}, \tag{9.1}$$

where $P_{\text{subthreshold}}$ is power associated with MOSFET subthreshold conduction, P_{leakage} is power associated with p-n junction leakage in the MOSFETs, P_{SC} is short-circuit dissipation, and P_{switch} is capacitance switching dissipation. In high-voltage CMOS circuits, the capacitance switching power is nearly always dominant so that

$$P \approx P_{switch} = \alpha f_{CLK} C_L V_{DD}^2, \tag{9.2}$$

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 lead 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, reduction of the supply voltage increases the propagation delays, which for symmetric circuits are

$$t_P = \frac{C_L}{K(V_{DD} - V_T)}\left[\frac{2V_T}{(V_{DD} - V_T)} + \ln\left(\frac{3V_{DD} - 4V_T}{V_{DD}}\right)\right] \approx \frac{1.6C_L}{K(V_{DD} - V_T)}, \tag{9.3}$$

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 of the supply voltage should be accompanied by reduction of the threshold voltages to maintain reasonable dynamic performance.

Figure 9.1 shows the propagation delay versus the supply voltage with the threshold voltage as a parameter, for a symmetric CMOS gate ($K = 100\ \mu A/V^2$)

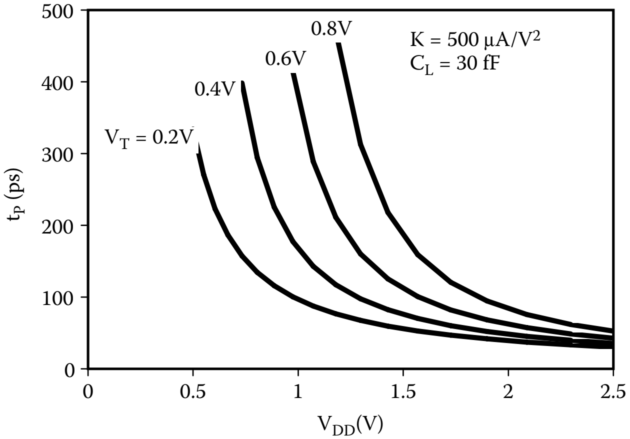


FIGURE 9.1
Propagation delay versus supply voltage with threshold voltage as a parameter, for a symmetric CMOS inverter with K = 500 μA/V² and loaded by 30 fF.

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, reduction of the threshold voltages increases the subthreshold currents and the associated dissipation. The subthreshold dissipation is

$$P_{subthreshold} \approx V_{DD}K(m-1)\left(\frac{kT}{q}\right)^2 \exp\left(\frac{-V_T}{mkT/q}\right), \tag{9.4}$$

where k is the Boltzmann constant, T is the absolute temperature, q is the electronic charge, and the subthreshold parameter m ranges between 1.1 and 2.0 for conventional MOSFETs operated at room temperature. This corresponds to a subthreshold swing between 65 and 120 mV. In fixed-threshold CMOS circuits, the absolute value of the threshold voltages must be at least 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.

9.3 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 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. Then circuits with more critical speed requirements can operate at higher voltages, whereas 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 hence 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 there is a tradeoff between power performance and complexity. Therefore, it is common to use two supply voltages: a high-voltage for output drivers and a low supply voltage for the internal circuitry.

Example 9.1 Sizing of CMOS Transistors Using Scaled V_{DD}

Suppose that a CMOS integrated circuit must drive 10 pF off-chip loads with a maximum propagation delay of 1 ns. Assuming 0.25 μm technology with t_{OX} = 8 nm, determine the minimum widths of the output driver transistors assuming $V_{TN} = |V_{TP}| = 0.3V$ and $V_{DD} = 3.3V$. Repeat for the case of $V_{DD} = 1.0V$.

Solution: First consider output drivers operating at 3.3 V. The delay factors are equal because of the symmetry in the threshold voltages:

$$\Gamma_P = \Gamma_N = \Gamma = \frac{1}{(V_{DD} - V_{TN})}\left[\frac{2V_{TN}}{(V_{DD} - V_{TN})} + \ln\left(\frac{3V_{DD} - 4V_{TN}}{V_{DD}}\right)\right]$$

$$= \frac{1}{(3.3V - 0.3V)}\left[\frac{2(0.3V)}{(3.3V - 0.3V)} + \ln\left(\frac{3(3.3V) - 4(0.3V)}{3.3V}\right)\right] = 0.390V^{-1}.$$

For both the p-MOS and n-MOS output drivers, the minimum device transconductance parameter is

$$K_{\min} \geq \frac{C_L\Gamma}{t_{P,\max}} = \frac{(10\times 10^{-12}F)(0.390V^{-1})}{(1\times 10^{-9}s)} = 3.9mA/V^2.$$

The process transconductance values for the p-MOS and n-MOS transistors are

$$k'_P = \frac{\mu_p\varepsilon_{OX}}{t_{OX}} = \frac{(230cm^2/Vs)(3.9)(8.85\times 10^{-14}F/cm)}{8\times 10^{-7}cm} = 100\mu A/V^2$$

and

$$k'_N = \frac{\mu_n\varepsilon_{OX}}{t_{OX}} = \frac{(580cm^2/Vs)(3.9)(8.85\times 10^{-14}F/cm)}{8\times 10^{-7}cm} = 250\mu A/V^2.$$

Therefore, the required device widths are

$$W_P \geq L_P \frac{K_{\min}}{k'_P} = (0.25\mu m)\frac{3900\mu A / V^2}{100\mu A / V^2} = 9.9\mu m$$

and

$$W_N \geq L_N \frac{K_{\min}}{k'_N} = (0.25\mu m)\frac{3900\mu A / V^2}{250\mu A / V^2} = 3.9\mu m \cdot$$

Using a 25% margin of safety, we could use device widths of $W_P = 12.4\mu m$ and $W_N = 4.9\mu m$.

For output drivers operating at 1.0 V, the delay factor is

$$\Gamma_P = \Gamma_N = \Gamma = \frac{1}{(V_{DD} - V_{TN})}\left[\frac{2V_{TN}}{(V_{DD} - V_{TN})} + \ln\left(\frac{3V_{DD} - 4V_{TN}}{V_{DD}}\right)\right]$$

$$= \frac{1}{(1.0V - 0.3V)}\left[\frac{2(0.3V)}{(1.0V - 0.3V)} + \ln\left(\frac{3(1.0V) - 4(0.3V)}{1.0V}\right)\right] = 2.064V^{-1}.$$

The minimum device transconductance parameter is

$$K_{\min} \geq \frac{C_L\Gamma}{t_{P,\max}} = \frac{(10\times10^{-12}F)(2.064V^{-1})}{(1\times10^{-9}s)} = 20.64mA / V^2 .$$

Therefore, the required device widths are

$$W_P \geq L_P \frac{K_{\min}}{k'_P} = (0.25\mu m)\frac{2.064\times10^{-2} A / V^2}{1\times10^{-4} A / V^2} = 51.6\mu m$$

and

$$W_N \geq L_N \frac{K_{\min}}{k'_N} = (0.25\mu m)\frac{2.064\times10^{-2} A / V^2}{2.5\times10^{-4} A / V^2} = 20.6\mu m \cdot$$

If we adopt a 25% safety margin, $W_P = 65\mu m$ and $W_N = 26\mu m$; therefore, the 1.0 V output drivers take up about five times as much silicon area as the 3.3 V output drivers having the same delay times.

In dual-voltage CMOS, the internal (low-voltage) circuitry can be optimized for low power, whereas the (high-voltage) output drivers can operate at higher voltage to conserve die area. All circuits are designed to just meet the speed requirements, to avoid wasting power or die area. Because the input/output circuits operate at a higher voltage than the internal circuitry, it is necessary to introduce voltage level shifters as shown in Figure 9.2.

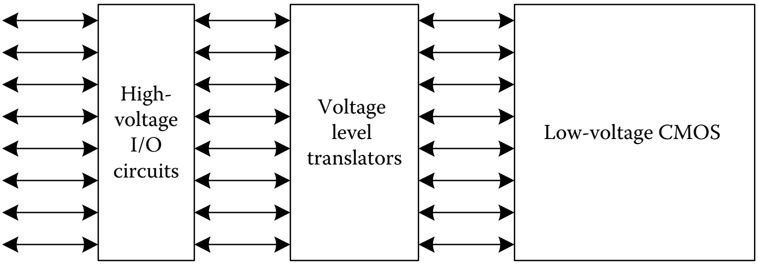


FIGURE 9.2
Dual-voltage CMOS.

9.4 Dynamic Voltage Scaling

The choice of supply voltage always involves a tradeoff between speed and power, regardless of the choice of threshold voltage(s). 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 input/output 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. This is because the output drivers can operate at a high supply voltage, for high speed, whereas 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, both from circuit to circuit and over time. Therefore, the choice of any single supply voltage for the core of the circuitry will result in wasted power. DVS allows the supply voltage to be adjusted dynamically for each block of circuitry to minimize the dissipation.

There are two basic approaches to DVS. The first uses a finite number of discrete supply voltages (discrete V_{DD} scaling), and the second uses a continuous variation of the supply voltage (arbitrary V_{DD} scaling). In either case, the supply voltage is produced by a switching power supply, which is 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 both be adjusted as appropriate for the necessary throughput to optimize the dissipation. Such a realization is shown in Figure 9.3. Here, the continuously adjustable V_{DD} is provided by the switching power supply. This power supply feeds both 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

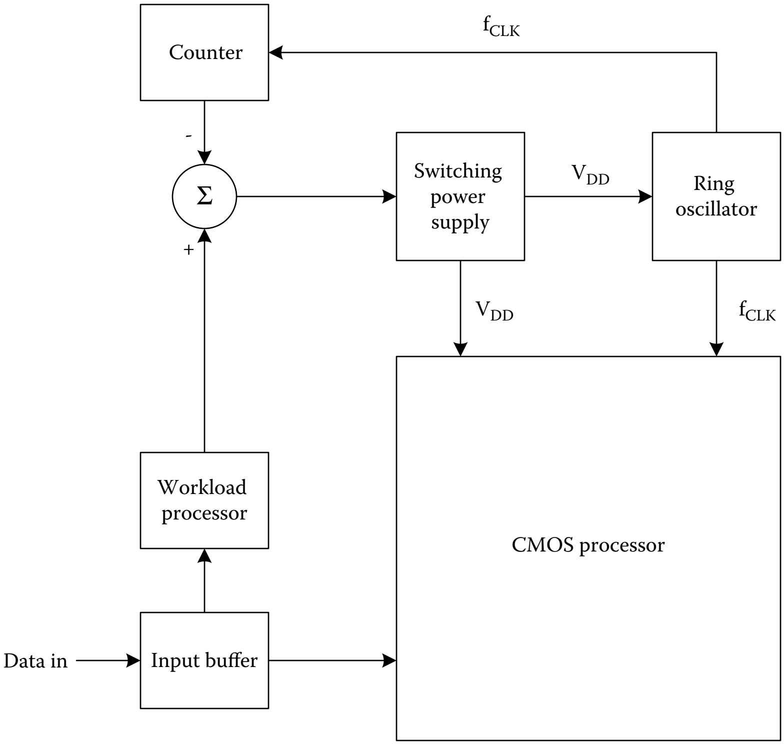


FIGURE 9.3
DVS system.

V_{DD} until it is just sufficient so 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, therefore 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.

9.5 Active Body Biasing

In low-voltage CMOS, choice of the threshold voltage involves a tradeoff between the speed and the subthreshold conduction. The active biasing scheme avoids this tradeoff because the threshold voltages of the transistors are adjusted dynamically. Therefore, circuits designed in this way are often called variable-threshold CMOS. To achieve threshold voltage adjustment, substrate bias control circuits drive the bodies of the n-MOSFETs and p-MOSFETs. Although the active biasing scheme consumes extra silicon area, to support the bias control circuitry, it is commonly used in modern microprocessors.

The basis for active biasing is the body effect. If there is a non-zero-bias V_{BS} applied between the source and the body of an n-MOSFET, then the threshold voltage is modified to

$$V_T = V_{TO} + \gamma\left(\sqrt{|V_{BS} + 2\phi_F|} - \sqrt{|2\phi_F|}\right), \tag{9.5}$$

where V_{TO} is zero-bias threshold, $2\phi_F$ is voltage across the semiconductor necessary to create a conducting channel (inversion layer), and γ is 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 the application of a positive bias on the body.

A general active biasing scheme takes the form shown in Figure 9.4. 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, whereas 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. This increases both V_{TN} and $|V_{TP}|$, thereby reducing the subthreshold conduction.

The use of active biasing relaxes the tradeoff between speed and subthreshold conduction. Therefore, the nominal threshold voltages can be much lower than those used in fixed-threshold circuits, and this allows a reduction of the supply voltage for low-power operation.

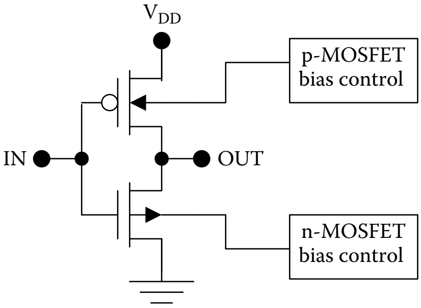


FIGURE 9.4
CMOS inverter with active body biasing.

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 additional reduction in the nominal threshold voltages without adverse consequences.

Example 9.2 Body Effect in Variable Threshold CMOS Circuits

Consider variable threshold CMOS with t_{OX} = 9 nm. Both 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 to ±0.3 V.

Solution: The body effect coefficient for the n-channel MOSFETs is

$$\gamma_N = \frac{\sqrt{2q\varepsilon_{Si}N_A}}{C_{ox}}$$

$$= \frac{\sqrt{2\left(1.602\times10^{-19}C\right)(11.9)\left(8.85\times10^{-14}F/cm\right)\left(10^{16}cm^{-3}\right)}}{(3.9)\left(8.85\times10^{-14}F/cm\right)/9\times10^{-7}cm}$$

$$= 0.151V^{1/2}$$

and

$$\phi_{FN} = \frac{kT}{q}\ln\left(\frac{n_i}{N_A}\right) = (0.026V)\ln\left(\frac{1.45\times10^{10}cm^{-3}}{10^{16}cm^{-3}}\right) = -0.35V \ .$$

The necessary body bias is

$$V_{BSN} = -\left[\frac{\Delta V_T}{\gamma_N} + \sqrt{|2\phi_{FN}|}\right]^2 - 2\phi_{FN}$$

$$= -\left[\frac{0.2V}{0.151V} + \sqrt{0.70V}\right]^2 + 0.70V = -4.0V \ .$$

For the p-channel MOSFETs,

$$\gamma_P = -\frac{\sqrt{2q\varepsilon_{Si}N_D}}{C_{ox}}$$

$$= -\frac{\sqrt{2\left(1.602\times10^{-19}C\right)(11.9)\left(8.85\times10^{-14}F/cm\right)\left(10^{16}cm^{-3}\right)}}{(3.9)\left(8.85\times10^{-14}F/cm\right)/9\times10^{-7}cm}$$

$$= -0.151V^{1/2}$$

and

$$\phi_{FP} = \frac{kT}{q}\ln\left(\frac{N_D}{n_i}\right) = (0.026V)\ln\left(\frac{10^{16}cm^{-3}}{1.45\times 10^{16}cm^{-3}}\right) = 0.35V.$$

The necessary body bias is

$$V_{BSP} = \left[\frac{\Delta V_T}{\gamma_P} + \sqrt{|2\phi_{FP}|}\right]^2 - 2\phi_{FP}$$

$$= \left[\frac{-0.2V}{-0.151V} + \sqrt{0.70V}\right]^2 - 0.70V = +4.0V \quad .$$

This example shows that relatively large body bias voltages may be required.

Example 9.3 Standby Power in Variable Threshold CMOS Circuits

Consider variable threshold CMOS with $K = 500\mu A/V^2$ and $V_{DD} = 1.0V$. Calculate the standby power as a function of V_T, assuming that the subthreshold conduction is dominant. $T = 300K$ and $m = 1.6$.

Solution: The standby power is given by

$$P_{subthreshold} \approx V_{DD}K(m-1)\left(\frac{kT}{q}\right)^2 \exp\left(\frac{-V_T}{mkT/q}\right)$$

$$= (1.0V)(500\times 10^{-6}A/V^2)(1.6-1)(0.026V)^2 \exp\left(\frac{-V_T}{1.6(26mV)}\right)$$

$$= (2.02\times 10^{-7}W)\exp\left(\frac{-V_T}{41.6mV}\right).$$

The results are plotted in Figure 9.5.

9.6 Multiple-Threshold CMOS

Multiple-threshold CMOS circuits can also be used to overcome the tradeoff between speed and subthreshold conduction inherent in single-threshold CMOS. Two or more distinct threshold voltages have been used in commercial microprocessors, but the simplest multiple threshold scheme is dual-threshold CMOS, in which the logic circuits are disconnected from the supply rails during standby operation. Therefore, these logic circuits can use

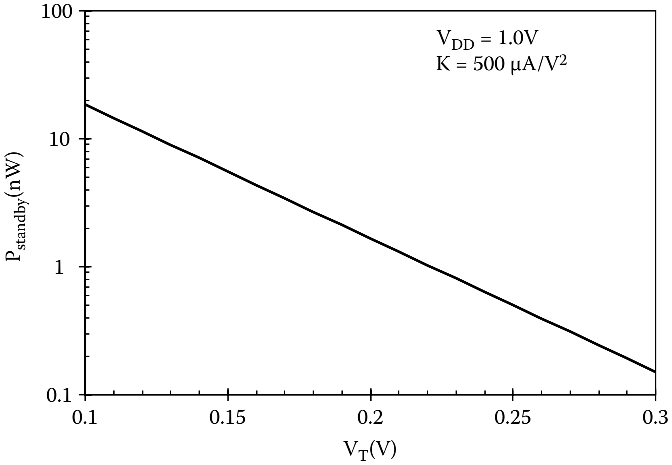


FIGURE 9.5
Standby dissipation versus the threshold voltage for variable voltage CMOS with V_{DD} = 1.0V and 500 × 10^{-6} A/V².

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.

Figure 9.6 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 ground (GND) line to the power rails. In the standby state, the enable signal is brought low. Both power switching transistors are cutoff 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 dual-threshold CMOS circuits reduces the standby dissipation (compared with the case of low-threshold CMOS) by a factor

$$\frac{P_{DTCMOS}}{P_{LTCMOS}} = \frac{K_H}{nK_L} 10^{-(V_{TH}-V_{TL})/S}, \tag{9.6}$$

where S is subthreshold swing, n is the number of low-threshold logic gates served by one pair of power switching transistors, KH is device transconductance parameter for high-threshold MOSFETs, KL is device

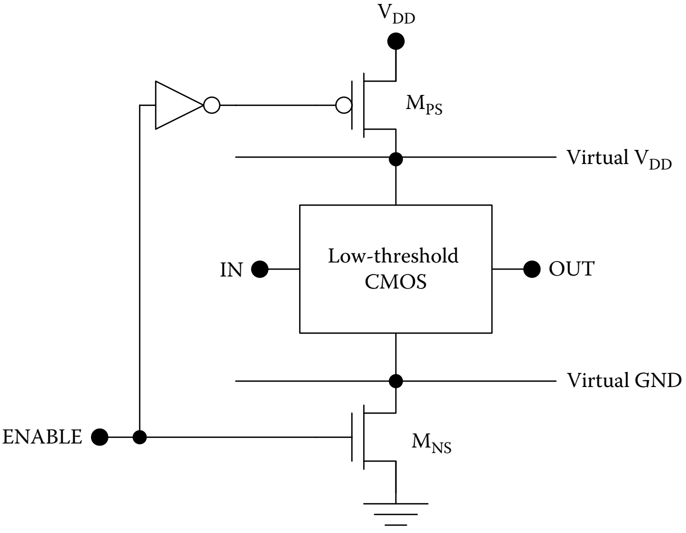


FIGURE 9.6
Dual-threshold CMOS.

transconductance for low-threshold MOSFETs, *VTH* is threshold voltage for high-threshold MOSFETS, and *VTL* is threshold voltage for low-threshold MOSFET. An issue with dual-threshold CMOS is that the circuits lose data when they are disconnected from the supply rails. To address this, latches (called *balloon circuits*) are used to retain data in the "sleep" state.

Example 9.4 Standby Power in Dual-Threshold CMOS Circuits

Estimate the reduction in the standby power associated with the use of dual-threshold CMOS with $V_{TH} = 0.3V$ and $V_{TL} = 0.1V$.

Solution: It is assumed that the power switching transistors are scaled sufficiently so that they only increase the propagation delays of the low-threshold circuits 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 with the use of low-threshold CMOS is approximately

$$\frac{P_{DTCMOS}}{P_{LTCMOS}} = \frac{K_H}{nK_L} 10^{-(V_{TH}-V_{TL})/S} = (10)10^{-(0.3V-0.1V)/0.1V} = 0.1,$$

corresponding to a 90% saving in standby dissipation.

9.7 Adiabatic Logic

Adiabatic logic circuits [45–47] conserve power by recycling electrical charge, whereas a conventional CMOS circuit uses each electrical charge only once. To see why this is so, consider the conventional CMOS circuit as shown in Figures 9.7 and 9.8. 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 9.7, the load capacitance charges to V_{DD} and the energy $C_LV_{DD}^2/2$ 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 9.8, 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 $C_LV_{DD}^2$ 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, thereby reducing the average current draw 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 Figures 9.9 and 9.10. In Figure 9.9, 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 R_{DP}.

If the initial voltage on the load capacitance is zero, then during the charge-up process, the voltage on the capacitor is given by

$$V_C(t) = \frac{I_{SOURCE}}{C_L} t\,. \tag{9.7}$$

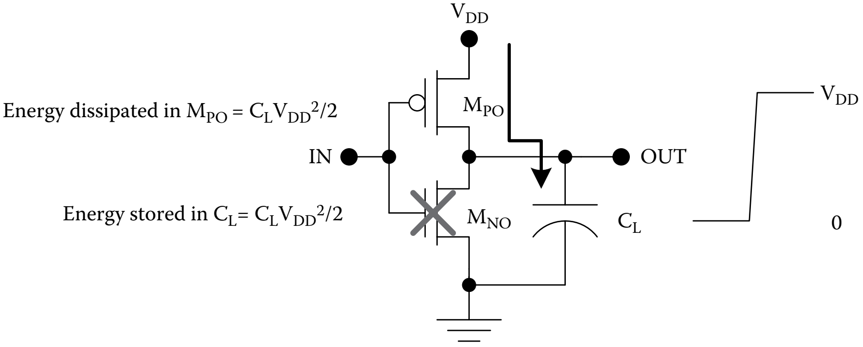

V_DD
Energy dissipated in M_PO = C_LV_DD^2/2
M_PO
IN
OUT
Energy stored in C_L= C_LV_DD^2/2
M_NO
C_L
V_DD
0

FIGURE 9.7
Charging of the load capacitance in conventional CMOS.

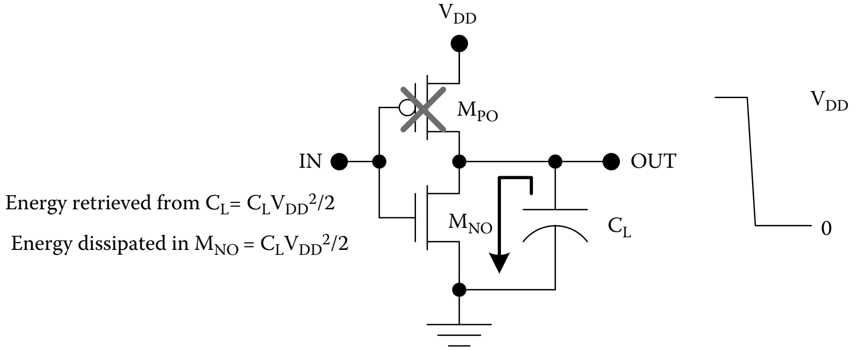

V_{DD}
V_{DD}
M_{PO}
IN
OUT
Energy retrieved from $C_L = C_L V_{DD}^2/2$
Energy dissipated in $M_{NO} = C_L V_{DD}^2/2$
M_{NO}
C_L
V_{DD}
0

FIGURE 9.8
Discharge of the load capacitance in conventional CMOS.

The time required to charge the load up to V_{DD} is therefore

$$t_R = \frac{V_{DD}C_L}{I_{SOURCE}}. \tag{9.8}$$

During the charge-up process, the dissipation in the p-MOSFET is constant:

$$P_R = R_{DP}I_{SOURCE}^2. \tag{9.9}$$

Therefore, the energy dissipated in the p-MOSFET during the entire process of charging the capacitor from zero to V_{DD} is

$$J_R = V_{DD}C_L R_{DP} I_{SOURCE}. \tag{9.10}$$

Therefore, the energy wasted in the p-MOSFET may be made arbitrarily small if the rise time is made arbitrarily long.

The use of a current source to discharge the load capacitance provides a similar benefit, as can be shown with the aid of Figure 9.10. If the initial

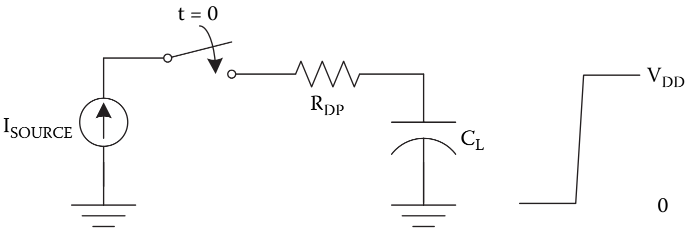


FIGURE 9.9
The charge-up of a load capacitance using a current source.

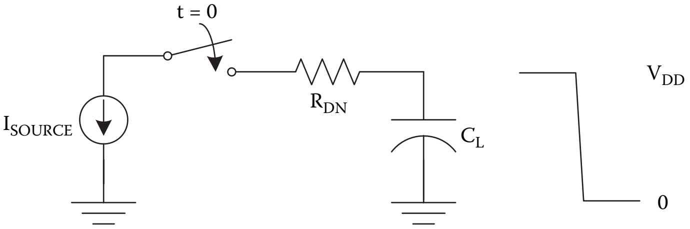


FIGURE 9.10
The discharge of the load capacitance using a current source.

voltage on the load capacitance is V_{DD}, then during the discharge process, the voltage on the capacitor is given by

$$V_C(t) = V_{DD} - \frac{I_{SOURCE}}{C_L} t \,. \tag{9.11}$$

The time required to discharge the load fully is therefore

$$t_F = \frac{V_{DD} C_L}{I_{SOURCE}} \,. \tag{9.12}$$

During the discharge process, the dissipation in the n-MOSFET is constant:

$$P_F = R_{DN} I_{SOURCE}^2 \,. \tag{9.13}$$

Therefore, the energy dissipated in the n-MOSFET during the entire process of discharging the load capacitor from V_{DD} to zero is

$$J_F = V_{DD} C_L R_{DN} I_{SOURCE} \,. \tag{9.14}$$

Therefore, 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 we compare the total energy wasted per switching cycle in the adiabatic circuit with that in the conventional CMOS case, assuming symmetric circuits with equal fall and rise times $(t_F = t_R = \tau)$, then

$$\frac{J_{Adiabatic}}{J_{Conventional}} = \frac{2V_{DD} C_L R_D I_{SOURCE}}{V_{DD}^2 C_L} = \frac{2R_D C_L}{\tau} \,. \tag{9.15}$$

In the adiabatic limit, this ratio will be zero. In real circuits, which dissipate a finite amount of power, there is a tradeoff between the dissipation and the speed.

In practical adiabatic circuits*, the current sources are implemented approximately using linear voltage ramps. Such a circuit is shown in Figure 9.11, which is an adiabatic inverter/buffer. With a logic zero input, the top transmission gate turns on but the bottom transmission gate stays off. During the ramp up of V_{RAMP}, the load capacitance C_{LN} at the non-inverting output charges with an approximately constant current, drawing energy from the ramp supply. The inverting 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 non-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. Hence, 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 9.12. 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 conventional static CMOS circuits. This is because complementary signals must be provided to drive the fan-out gates. Therefore, as with many of the other low-power CMOS strategies, there is a tradeoff between dissipation and circuit area.

9.8 Silicon-on-Insulator

SOI 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 because of the improved subthreshold characteristics of SOI transistors compared with bulk silicon MOSFETs.

Over the years, a number of SOI fabrication technologies have emerged. For example, silicon-on-sapphire has been explored extensively for power

* 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.

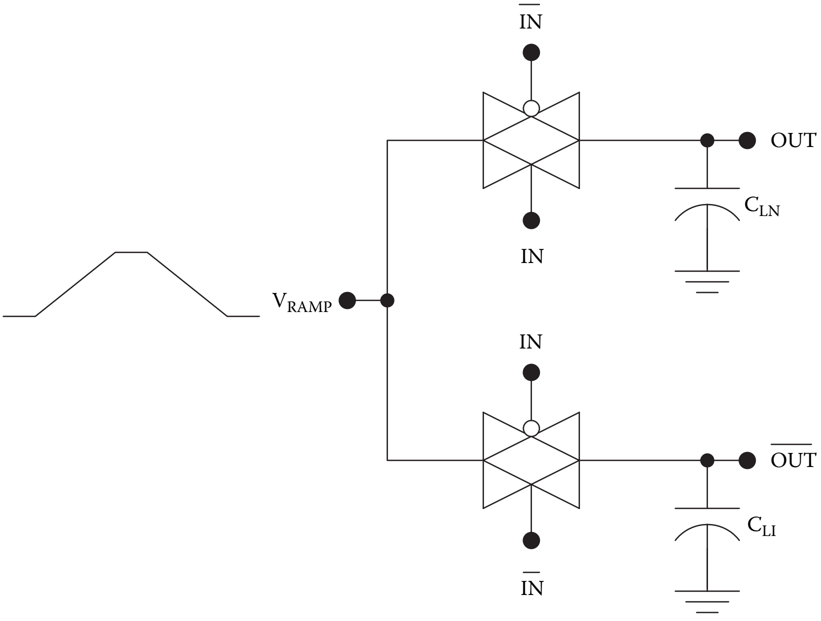


FIGURE 9.11
Adiabatic CMOS inverter/buffer.

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.

In the following sections, we will review two SOI fabrication technologies: separation by implantation of oxygen (SIMOX) and wafer bonding (WB). Then we will delineate the features of fully depleted (FD) and partially depleted (PD) SOI transistors. Finally, we will consider the application of SOI to low-power CMOS.

9.8.1 SOI Technologies: SIMOX and Wafer Bonding

At the present time, there are two basic fabrication technologies for SOI: SIMOX [32, 33] and 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 BOX. The WB process involves the bonding of an oxidized device wafer (DW) to a handle wafer (HW), followed by removal of all but a thin layer from the DW. Two important variations of the WB process are the epitaxial layer transfer (ELTRAN)

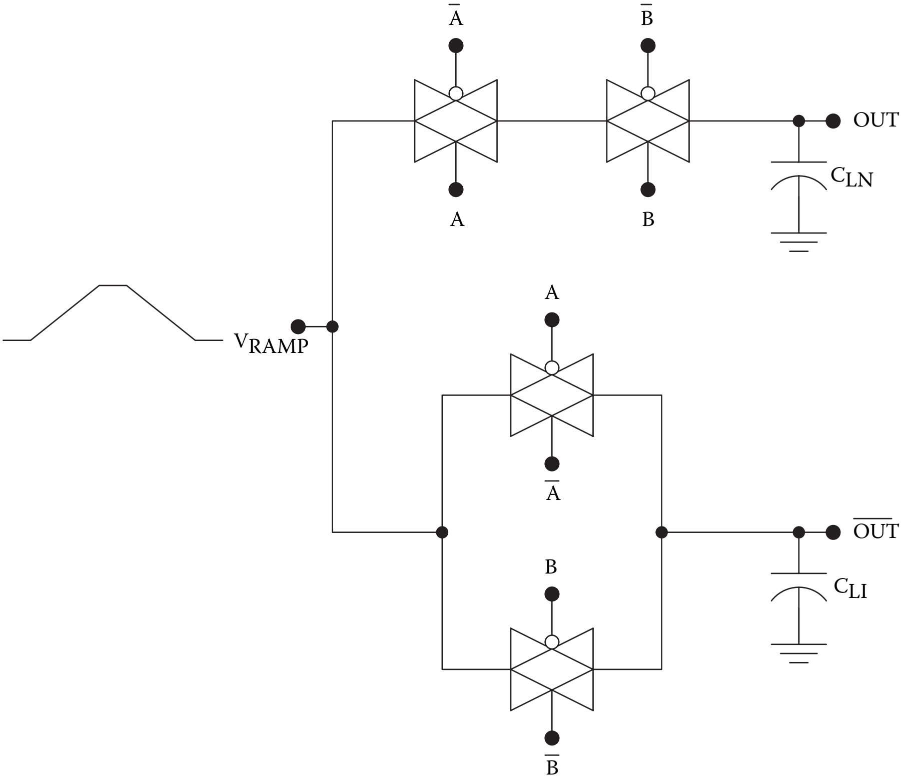


FIGURE 9.12
Adiabatic CMOS AND2/NAND2 gate.

process and the UNIBOND™ process. Both 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 Figures 9.13 through 9.15 . First, a low dose $\left(4\times10^{17}cm^{-2}\right)$ of oxygen ions is implanted at an energy of about 200 keV (Figure 9.13). Next, the wafer is treated at a high temperature (>1300°C) to anneal out the defects created by the ion implantation process (Figure 9.14). Finally, the wafer is subjected to an internal thermal oxidation (ITOX) process, which increases the thickness of the BOX layer to a usable value (Figure 9.15).

Typically, the resulting SOI layer is 50–100 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 Figures 9.16 through 9.21. In contrast to the SIMOX process, WB processes such as ELTRAN require two wafers. These are called the DW and the HW. First, the DW is treated by anodization

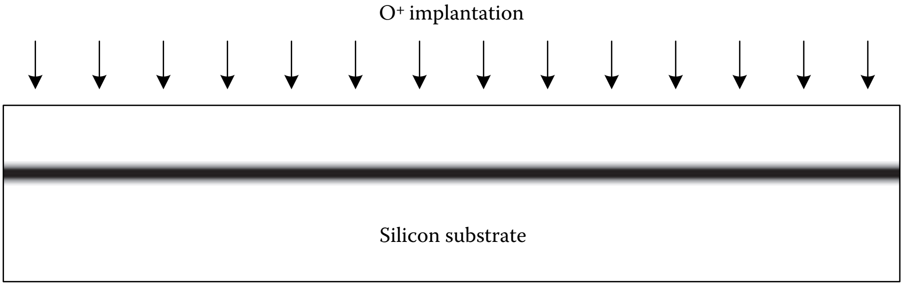


FIGURE 9.13
SIMOX process sequence A. The wafer is implanted with a low dose (4×10^{17} cm^{-2}) of oxygen at an energy of approximately 200 keV.

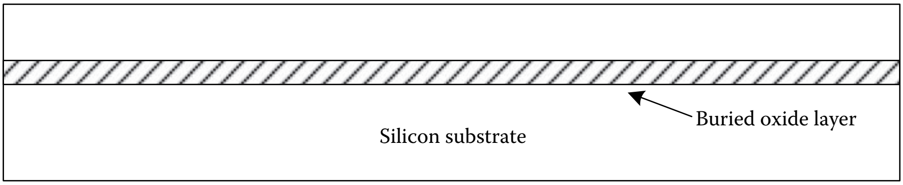


FIGURE 9.14
SIMOX process sequence B. Ion implantation damage is annealed out at >1300°C.

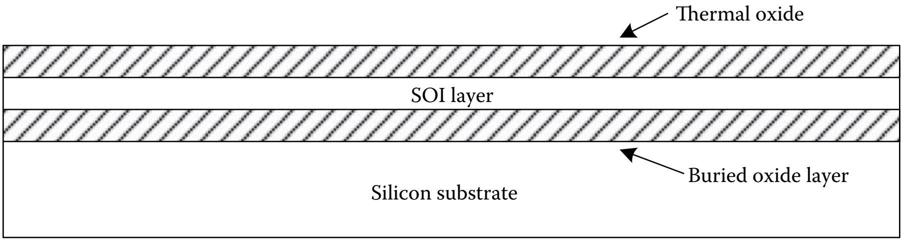


FIGURE 9.15
SIMOX process sequence C. ITOX at >1300°C is used to increase the thickness of the BOX layer.

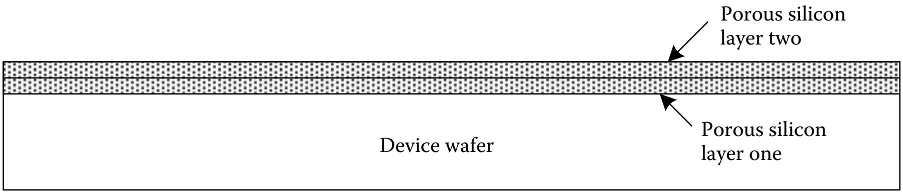


FIGURE 9.16
ELTRAN process sequence A. The DW is anodized in a two-step process to result in two layers of porous silicon having distinctly different porosity.

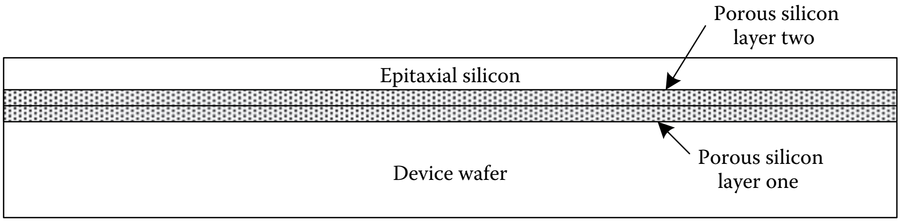


FIGURE 9.17
ELTRAN process sequence B. An epitaxial layer of device-quality silicon is grown on the porous silicon by vapor phase epitaxy.

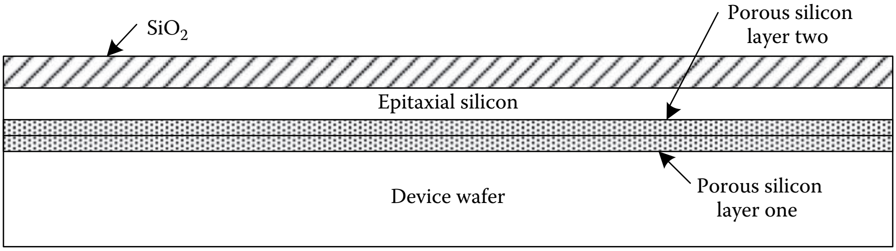


FIGURE 9.18
ELTRAN process sequence C. Thermal oxidation results in an oxide layer on top of the epitaxial layer of silicon.

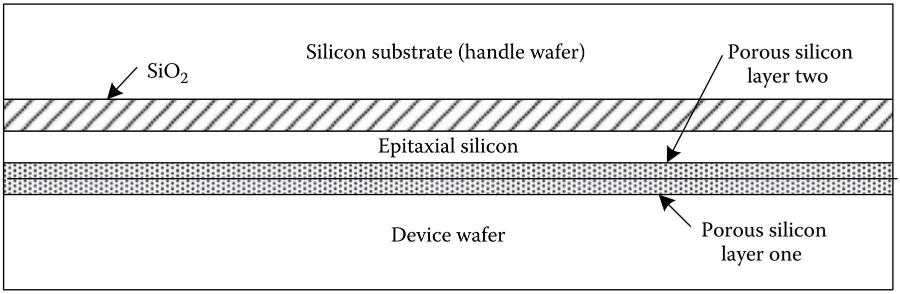


FIGURE 9.19
ELTRAN process sequence D. The HW is bonded to the processed DW.

to create a layer of porous silicon on its surface (Figure 9.16). The anodization process is designed to result in two distinct layers with different porosity. Second, a high-quality epitaxial layer of silicon is grown on top of the porous silicon by vapor phase epitaxy (Figure 9.17). Third, thermal oxidation is used to create the oxide layer on top of the epitaxial device layer (Figure 9.18). Fourth, an HW is bonded to the processed DW (Figure 9.19). Fifth, a water jet is used to separate the structure between the two distinct layers of porous

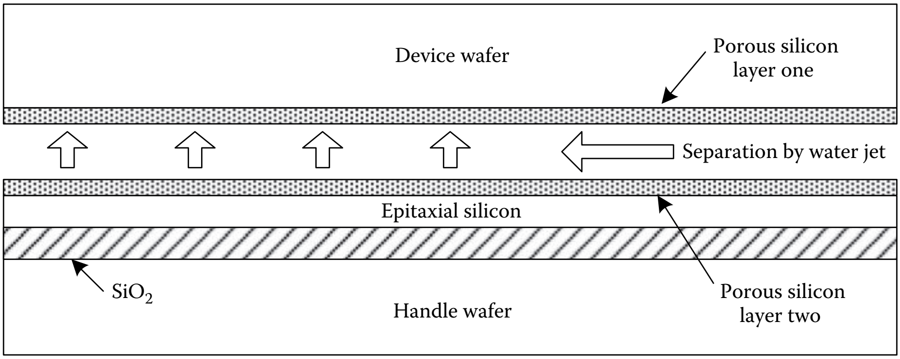


FIGURE 9.20
ELTRAN process sequence E. The HWs and DWs are separated between the distinct porous silicon layers by a water jet.

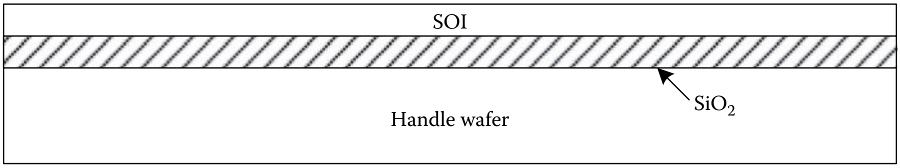


FIGURE 9.21
ELTRAN process sequence F. Chemical etching is used to remove the remaining porous silicon from the SOI layer.

silicon (Figure 9.20). Finally, a chemical etch is used to remove the remaining porous silicon (Figure 9.21).

Another WB approach is the UNIBOND^{TM} process, illustrated in Figures 9.22 through 9.24. First, the DW is thermally oxidized. Then hydrogen ions are implanted through this oxide layer with a dose of about 10^{16} cm^{-2} (Figure 9.22). Next, the HW is bonded to the DW (Figure 9.23). Then, an annealing process at 400–600°C serves to split the wafers apart, and the resulting SOI wafer is chemically polished (Figure 9.24).

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.

9.8.2 SOI MOSFETs: Fully Depleted or Partially Depleted

Depending on the thickness of the SOI layer, the bodies of the resulting MOSFETs may be PD [38–40] or FD [41–44]. FD and PD SOI n-MOSFETs are depicted in Figure 9.25. Typically, FD MOSFETs are fabricated with an SOI

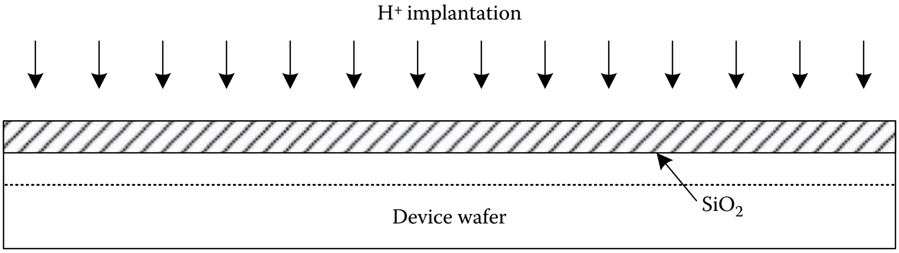


FIGURE 9.22
UNIBOND™ process sequence A. The DW is thermally oxidized and then implanted with hydrogen ions through the oxide.

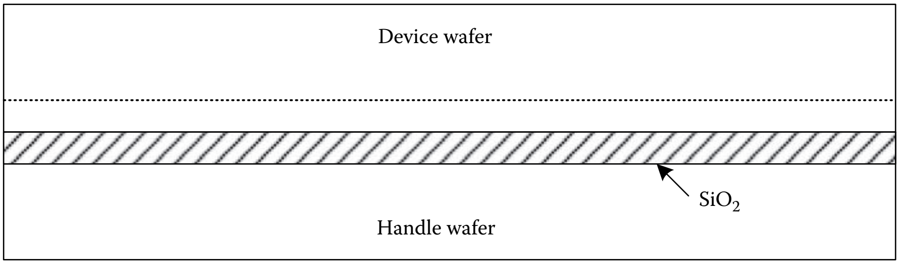


FIGURE 9.23
UNIBOND™ process sequence B. The DW is bonded to the HW.

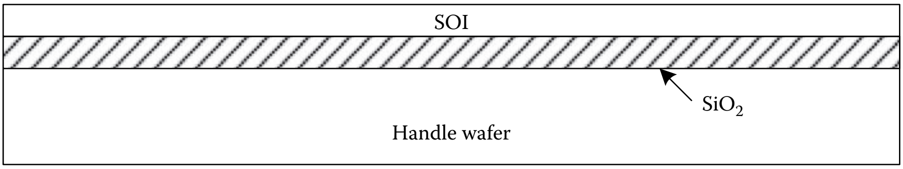


FIGURE 9.24
UNIBOND™ process sequence C. An annealing process at 400–600ºC splits the wafers apart, and the SOI wafer is chemically polished.

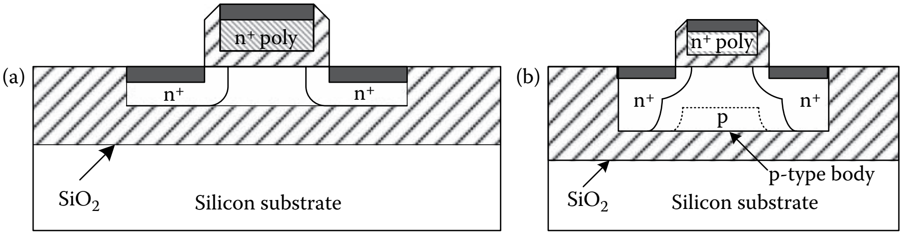


FIGURE 9.25
(a) FD SOI n-MOSFET and (b) PD SOI n-MOSFET.

layer approximately 50 nm thick. Therefore, with zero bias, the p-type silicon region under the channel becomes FD by the built-in potentials at the S/D 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. Both PD and FD SOI MOSFETs have particular advantages, and as such both device types have been applied in commercial microprocessors since 1999.

PD 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 used with PD SOI MOSFETs, to either reduce subthreshold conduction or compensate for processing variations in the threshold voltages. This is not possible with FD devices, because of the absence of a p-type body region. On the other hand, FD transistors exhibit near-ideal (~65 mV at room temperature) values for the subthreshold slope; as a consequence, 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.

PD devices can be made with floating bodies (without body contacts). These devices exhibit a *floating body effect*, whereby holes created by impact ionization near the drain accumulate in the body region. This creates a self body bias that lowers the threshold voltage for the device. This can be advantageous in some logic circuits, in which it increases the speed. In pass transistor applications such as DRAMs, however, the floating body effect is problematic because the reduction in the threshold voltage increases the off-state leakage.

9.8.3 SOI for Low-Power CMOS

SOI is an important technology for low-power CMOS because SOI MOSFETs have superior subthreshold characteristics compared with bulk silicon MOSFETs. In FD SOI MOSFETs, the subthreshold swing is near the ideal value of 60 mV at room temperature. This combined with reduced process-induced variations in the threshold voltage allow the implementation of SOI CMOS scaled down to $V_{DD} = 0.4V$ and $V_T = 0.1V$. This reduction in the supply voltage decreases the capacitance switching power to one-sixth of the value for bulk CMOS designed with $V_{DD} = 1.0V$ and $V_T = 0.3V$. Additional 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

$$S = 2.3\frac{mkT}{q}, \tag{9.16}$$

where k is the Boltzmann constant, T is the absolute temperature, q is the electronic charge, and

$$m = 1 + \frac{C_{dm}}{C_{ox}}, \tag{9.17}$$

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 FD SOI MOSFETS such as those shown in Figure 9.26, 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, there is no neutral n-type region under the channel of the p-MOSFET, although it is fabricated in an n-type well. Therefore,

$$C_{dm} \rightarrow 0, \tag{9.18}$$

and

$$m \rightarrow 1. \tag{9.19}$$

This results in near-ideal values of the subthreshold swing. Of course, the semiconductor depletion capacitance can never be identically zero, so the subthreshold swing parameter m will be greater than one and the subthreshold current will always be finite. At room temperature, the ideal subthreshold swing is

$$S \approx 2.3\frac{kT}{q} = 60mV \cdot \tag{9.20}$$

Experimentally measured values of the subthreshold swing in FD SOI MOSFETs are typically 65 mV, corresponding to $m \approx 1.1$.

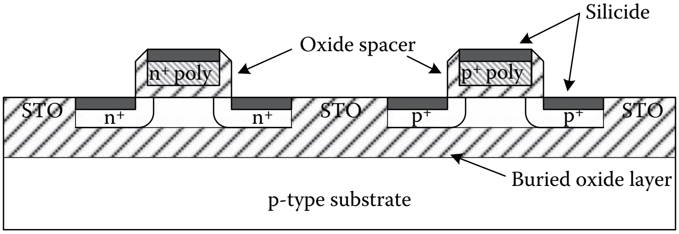


FIGURE 9.26
SOI CMOS transistors fabricated by the ITOX-SIMOX process.

Example 9.6 Standby Dissipation in SOI CMOS

Compare the standby dissipation for SOI CMOS $(S = 65mV)$ with that for bulk CMOS $(S = 100mV)$ at room temperature (300 K). Assume V_{DD} = 1V, $V_{TN} = |V_{TP}| = 0.1V$, and $K = 500\mu A / V^2$ 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

$$P_{subthreshold} \approx V_{DD}K(m-1)\left(\frac{kT}{q}\right)^2 \exp\left(\frac{-V_T}{mkT/q}\right)$$

$$= (1.0V)\left(500\times10^{-6}A/V^2\right)(1.6-1)(0.026V)^2 \exp\left(\frac{-0.1V}{41.6mV}\right) = 1.8nW\ .$$

For the FD SOI CMOS with S = 65 mV, the corresponding subthreshold parameter is $m = 1.09$, and the standby dissipation per gate is

$$P_{subthreshold} \approx V_{DD}K(m-1)\left(\frac{kT}{q}\right)^2 \exp\left(\frac{-V_T}{mkT/q}\right)$$

$$= (1.0V)\left(500\times10^{-6}A/V^2\right)(1.09-1)(0.026V)^2 \exp\left(\frac{-0.1V}{28.3mV}\right) = 0.9nW\ .$$

Therefore, the FD SOI CMOS reduces the standby power to a value roughly half of that for bulk CMOS, with all other things being equal.

9.9 Practical Perspective

For practical perspective articles, see the dynamic website at http://www.engr.uconn.edu/ece/books/ayers.

9.10 Summary

Low-power CMOS integrated circuits are important for portable, battery-operated systems. In these circuits, the capacitance switching power is often dominant so the dissipation per gate is proportional to the load capacitance times the square of the supply voltage. Aggressive scaling of the physical dimensions of the MOSFETs has lead to significant reductions in the load capacitances and the dissipation per gate, but the most effective way to reduce power is by scaling the supply voltage. However, this must be accompanied by threshold voltage reduction to maintain acceptable propagation delays. There is a limit to the scaling of threshold voltage imposed by the standby dissipation, however.

The choice of supply voltage always involves a tradeoff between speed and power, regardless of the choice of threshold voltage(s). 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 input/output circuits. This results in wasted power in other parts of the integrated circuit.

The use of two or more supply voltages in *multiple-voltage CMOS* alleviates this problem to a great extent, because the output drivers can operate at a higher supply voltage, for high speed, whereas 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, both from circuit to circuit and over time. Therefore, the choice of any single supply voltage for the core of the circuitry will result in wasted power. 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.

Active body biasing avoids the tradeoff 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, whereby 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, whereas inactive circuits are biased to have increased threshold voltages reduced subthreshold conduction.

Multiple-threshold CMOS circuits can also be used to overcome the tradeoff 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, 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 that the standby power can be greatly reduced by this approach.

SOI is an important technology for low-power CMOS because SOI MOSFETs have superior subthreshold characteristics compared with bulk silicon MOSFETs. At the present time, SOI wafers are produced by the SIMOX and WB approaches. Both methods use the same wafers as bulk CMOS circuits, which are available in large diameter, with high quality, and at low cost. SOI MOSFETs may be PD or FD. In FD SOI MOSFETs, the subthreshold slope is near the ideal value of 60 mV at room temperature. This combined with reduced process-induced variations in the threshold voltage allow the implementation of SOI CMOS scaled down to $V_{DD} = 0.4V$ and $V_T = 0.1V$. This reduction in the supply voltage decreases the capacitance switching power to one-sixth of the value for bulk CMOS designed with $V_{DD} = 1.0V$ and $V_T = 0.3V$. Additional 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%.

Adiabatic logic circuits have been proposed to conserve power by recycling electrical charge, in contrast with conventional CMOS gates that use each electrical charges 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 $C_LV_{DD}^2/2$. 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 $\frac{1}{10}$ that of conventional CMOS.

9.11 Exercises

E9.1. Suppose that 1.0 V CMOS is fabricated using 0.25 μm technology and $t_{ox} = 6nm$. Consider inverters with $V_{TN} = |V_{TP}| = V_T$, $L_N = L_P = 0.25\mu m$, $W_N = 1.0\mu m$, $W_P = 2.5\mu m$, and $L_{OV} = 0.05\mu m$. (1) Calculate and plot the propagation delay (assuming five on-chip loads) versus the threshold voltage ($0.1V \le V_T \le 0.6V$). (2) Calculate and plot the standby dissipation versus the subthreshold voltage (0.1 V ≤ V_T ≤ 0.6V). Assume the subthreshold conduction is dominant and the subthreshold swing is 95 mV.

E9.2. Consider 0.25 μm CMOS technology with $t_{ox} = 6nm$. Dual supply voltages are to be used. $V_{TN} = |V_{TP}| = 0.3V$ for all devices. (1) Choose V_{DDL} such that symmetric inverters with five on-chip loads can switch with $t_P \le 250ps$. (2) Choose V_{DDH} such that output drivers with $(W_N + W_P) \le 50\mu m$ can drive 5 pF loads with $t_P \le 1ns$.

E9.3. Consider DVS implemented in 0.25 μm CMOS technology with $t_{ox} = 6nm$ and $L_{OV} = 0.05\mu m$. A ring oscillator is designed to mimic the critical path in the system using 13 stages of minimum size inverters with load capacitances at each stage equivalent to six on-chip loads. (1) Calculate and plot the ring oscillator frequency as a function of the supply voltage assuming $V_{TN} = |V_{TP}| = 0.3V$ for all devices. $1V \le V_{DD} \le 5V$. (2) Calculate and plot the switching power versus 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.

E9.4. Consider variable threshold CMOS fabricated with 0.25 μm technology. The n-MOSFETs and the p-MOSFETs both have 2×10^{16} cm^{-3} channel doping. $\pm$3.3 V is available for active body biasing of the MOSFETs. The standby threshold voltages (with body bias) are to be $\pm$0.3 V. What is the minimum value for the absolute value of the nominal threshold voltages (as fabricated, without body bias)?

E9.5. 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 $t_{ox} = 6nm$, $L_N = L_P = 0.25\mu m$, $W_N = 1.0\mu m$, $W_P = 2.5\mu m$, and $L_{OV} = 0.05\mu m$. Assume the inverters are loaded with three 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%.

E9.6. Show that, for an adiabatic logic gate that charges the load capacitance stepwise with n voltage steps, the capacitance switching dissipation is

$$P_{switch} = \frac{\alpha f_{CLK} V_{DD}^2 C_L}{n}.$$

For additional exercise problems, see the dynamic website at http://www.engr.uconn.edu/ece/books/ayers.

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, Digest of Technical 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 Electronics 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 Electronics 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, Digest of Technical 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, Digest of Technical 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 Electronics Design*, 201, 2001.
14. Sakurai, T., Kawaguchi, H., and Kuroda, T., Low-power CMOS design through VTH control and low-swing circuits**.** *Proc. 1997 Int. Symp. Low Power Electronics 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 Dev.*, 36, 1725, 1898.
22. Auberton-Herve, A.J., SOI: materials to systems. *1996 Int. Electron Dev. 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 Dev.*, 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 & scaling trends, Digest of Technical 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. Microelectronics*, 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., Takagi, S., Advanced SOI p-MOSFETs with strained-Si channel on SiGe-on-insulator substrate fabricated by SIMOX technology. *IEEE Trans. Electron Dev.*, 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., 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. Microelectronics,* 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 Electronics 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.

10

Bistable Circuits

10.1 Introduction

Bistable circuits exhibit two stable states that can represent logic 1 and logic 0. These include latches and flip-flops, 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. An example is the Schmitt trigger, which exhibits hysteresis and is useful in this regard.

There are two requirements for the realization of bistable operation. These 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 10.1. There are two stable states for this circuit. State 1 is

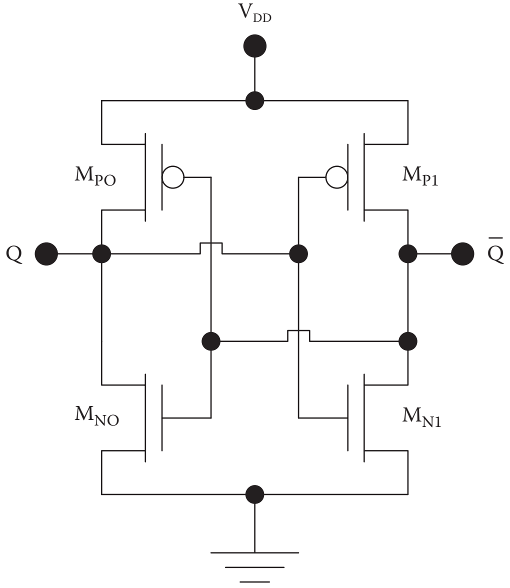


FIGURE 10.1
Cross-coupled CMOS inverters forming a bistable latch.

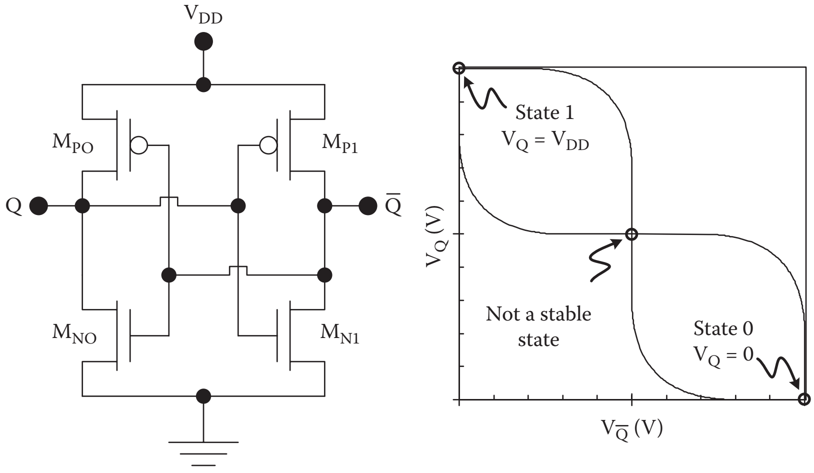


FIGURE 10.2
Cross-coupled CMOS inverters with their voltage transfer characteristics.

characterized by $Q = 1$ ($\overline{Q} = 0$), whereas state 0 is characterized by $Q = 0$ ($\overline{Q} = 1$). These two stable states are indicated in Figure 10.2 and represent two of the points of intersection on the *butterfly curve* showing the characteristics for the individual inverters. The third point of intersection at $V_Q = V_{\overline{Q}} = V_{DD}/2$ is *not* a stable state, so the circuit is bistable.

The circuit of Figure 10.1 has limited usefulness because it has no input connections and there is no way to actively set the circuit to state 0 or state 1. Instead, the state is set randomly during power up. In contrast, practical bistable circuits provide inputs so the state can be written as well as read.

In the following sections, we will consider latches, flip-flops, and Schmitt triggers [1–3]. 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 11.

Example 10.1 Asymmetric CMOS Latch

Determine the stable states for the CMOS latch made of mismatched inverters as shown in Figure 10.3.

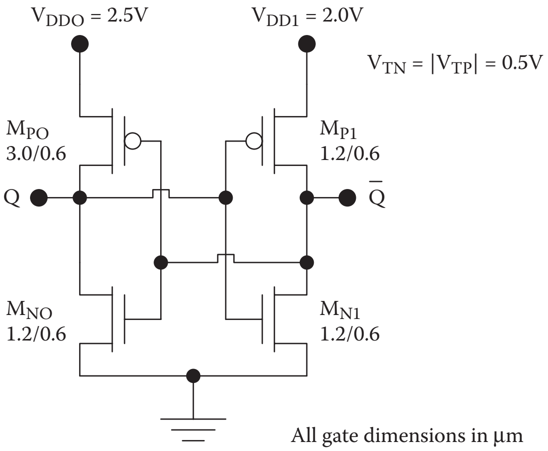


FIGURE 10.3
Example asymmetric CMOS latch.

Solution: For the inverter comprising M_{PO} and M_{NO}, $K_{NO}/K_{PO} = 1$ and

$$V_Q = \begin{cases} 2.5V; & (\text{regime } 1; \quad V_{IN} \le 0.5V) \\ V_{\bar{Q}} + 0.5V + \sqrt{\left(V_{\bar{Q}} - 2.0V\right)^2 - \left(V_{\bar{Q}} - 0.5V\right)^2}; & (\text{regime } 2; \quad 0.5V \le V_{IN} \le 1.25V) \\ 1.25V; & (\text{regime } 3; \quad V_{IN} \approx 1.25V) \\ V_{\bar{Q}} - 0.5V - \sqrt{\left(V_{\bar{Q}} - 0.5V\right)^2 - \left(V_{\bar{Q}} - 2.0V\right)^2}; & (\text{regime } 4; \quad 1.25V \le V_{IN} \le 2.0V) \\ 0; & (\text{regime } 5; \quad V_{IN} \ge 2.0V) \end{cases}$$

For the inverter on the right-hand side (comprising M_{P1} and M_{N1}), $K_{N1}/K_{P1}=3$ and $V_M = 0.87V$. Therefore,

$$V_{\bar{Q}} = \begin{cases} 2.0V; & (\text{regime } 1; \quad V_{IN} \le 0.5V) \\ V_Q + 0.5V + \sqrt{\left(V_Q - 1.5V\right)^2 - 3\left(V_Q - 0.5V\right)^2}; & (\text{regime } 2; \quad 0.5V \le V_{IN} \le 0.87V) \\ 1.00V; & (\text{regime } 3; \quad V_{IN} \approx 0.87V) \\ V_Q - 0.5V - \sqrt{\left(V_Q - 0.5V\right)^2 - \left(1/3\right)\left(V_Q - 1.5V\right)^2}; & (\text{regime } 4; \quad 0.87V \le V_{IN} \le 1.5V) \\ 0; & (\text{regime } 5; \quad V_{IN} \ge 1.5V) \end{cases}$$

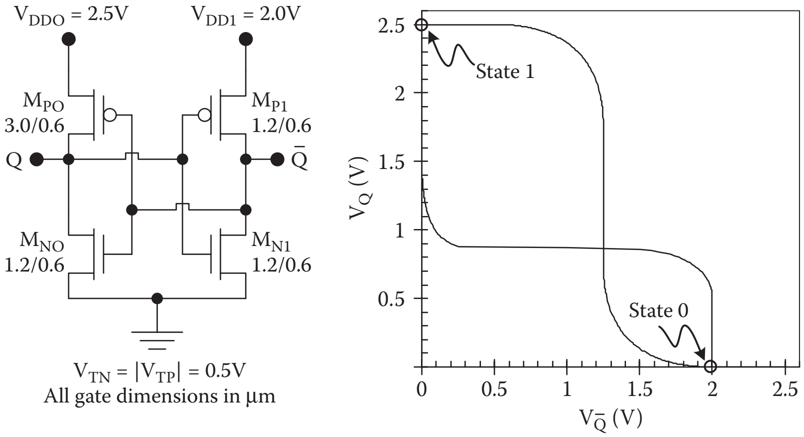


FIGURE 10.4
Asymmetric CMOS latch and butterfly curve.

The two characteristics are plotted in Figure 10.4, and their intersections at $\left(V_Q, V_{\bar{Q}}\right) = (2.5V, 0)$ and $\left(V_Q, V_{\bar{Q}}\right) = \left(0, 2.0V\right)$ represent the stable states.

10.2 Set-Reset Latch

A latch made from cross-coupled inverters does not allow active setting of the logic state, but this can be addressed by adding four input transistors as shown in Figure 10.5. If the input R is brought high, M_{NR} will be linear and M_{PR} will be cutoff. Therefore, the next state of Q will be "0" ($Q \to 0$), and R is therefore called the *reset* input. If the input S is brought high, M_{NS} will be linear, whereas M_{PS} will be cutoff. Therefore, $Q \to 1$, $\overline{Q} \to 0$, and S is referred to as the *set* input.

The CMOS set-reset (SR) latch illustrated in Figure 10.5 actually comprises two cross-coupled NOR gates and is sometimes called an NOR latch. The logic representation and truth table for the NOR SR latch are shown in Figure 10.6.

An RS latch may also be constructed using two cross-coupled NAND gates, as shown in Figure 10.7. This circuit differs from the NOR RS latch in that 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 Figures 10.8 and 10.9.

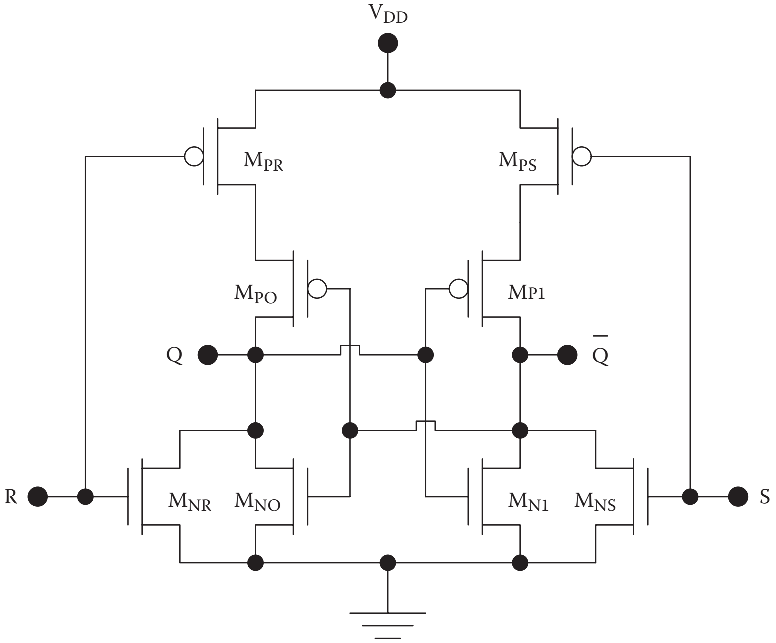
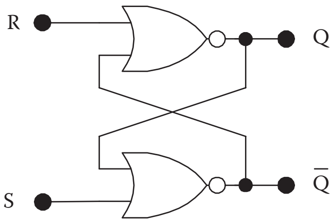
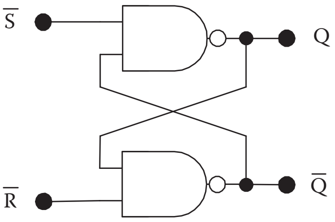


FIGURE 10.5
CMOS SR latch.



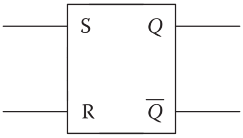
R	S	Q_{N+1}
0	0	Q_N
0	1	1
1	0	0
1	1	Not used

FIGURE 10.6
NOR SR latch and truth table.



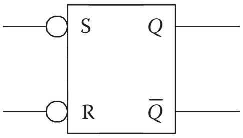
$\overline{R}$	$\overline{S}$	Q_{N+1}
0	0	Not used
0	1	1
1	0	0
1	1	Q_N

FIGURE 10.7
NAND SR latch and truth table.



R	S	Q_{N+1}
0	0	Q_N
0	1	1
1	0	0
1	1	Not used

FIGURE 10.8
NOR SR latch symbol 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 10.9
NAND SR latch symbol and truth table.

10.3 SR Flip-Flop

With RS latches, there is always an ambiguous input condition (both S and R active) that must be avoided. This situation can be alleviated somewhat by clocking the circuit as shown in Figure 10.10. In the clocked circuit, the S and R inputs are inactive unless the clock signal is high. As long as S and R 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*.

10.4 JK Flip-Flops

The ambiguous input condition of the SR latches and flip-flop can be avoided altogether by the use of feedback as shown in Figure 10.11. The resulting device is called a JK flip-flop, with the logic symbol and truth table as shown in Figure 10.12.

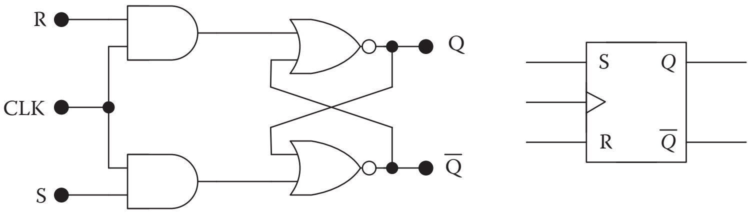


FIGURE 10.10
Clocked SR flip-flop.

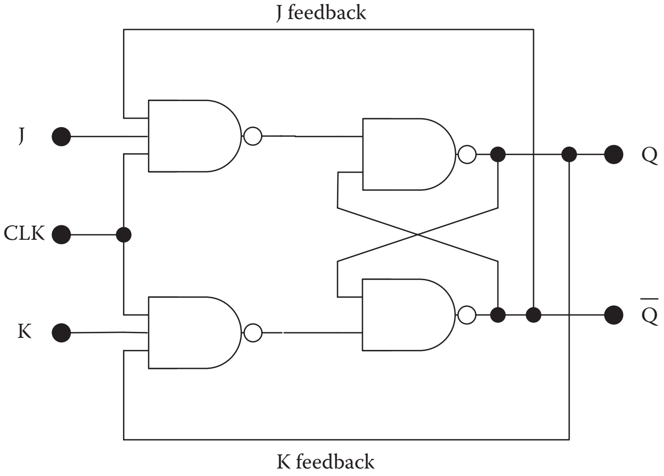
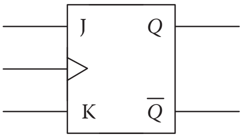


FIGURE 10.11
JK flip-flop.



J	*K*	Q_{N+1}
0	0	Q_N
0	1	1
1	0	0
1	1	$\overline{Q_N}$

FIGURE 10.12
JK flip-flop symbol and truth table.

In the JK flip-flop, the J input is active only when $(CLK, \overline{Q}) = (1,1)$, but 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 10.13. The premise for this figure is that logic 1 is applied to both J and K, and a clock pulse of long duration is applied at t = 0. After two propagation delays, $\overline{Q}$ goes high, and this begins 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. This is undesirable because the final output state will depend on the 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 two propagation delays:

$$2t_P < \frac{T}{2} < 4t_P . \tag{10.1}$$

Fortunately, the restriction on the duration of the clock pulse can be removed by either the use of a master-slave design or edge triggering.

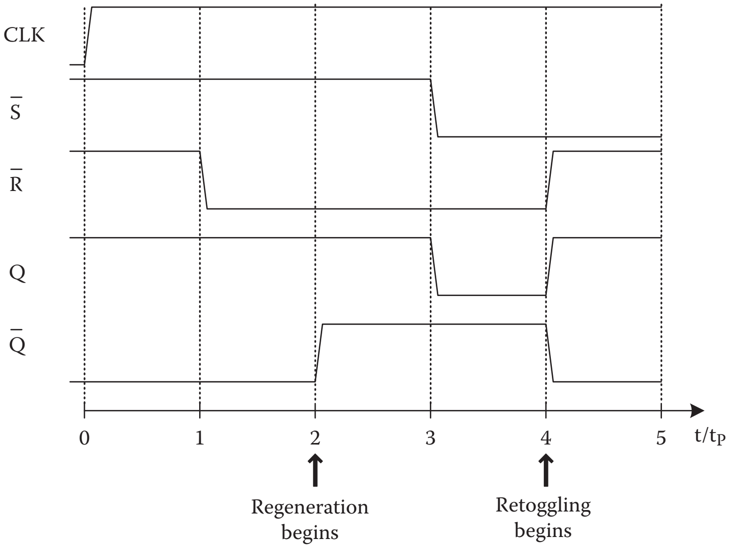


FIGURE 10.13
Timing diagram for a JK flip-flop.

The master-slave design involves the use of two cascaded JK flip-flops as shown in Figure 10.14. 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 therefore 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 high, the slave switches based only on the output of the master. As a result of this 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. Figure 10.15 shows the circuit diagram for a CMOS master-slave JK flip-flop. It contains 38 MOS transistors: eight each for two SR latches, six each for two NAND3 gates, four each for two NAND2 gates, and two for the inverter.

The edge-triggered flip-flop exploits the differences in propagation delays for different paths within the circuits. One such design is shown in Figure 10.16. During 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 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

$$t_{active} = t_{P1} + t_{P2} - t_{P3} . \tag{10.2}$$

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.

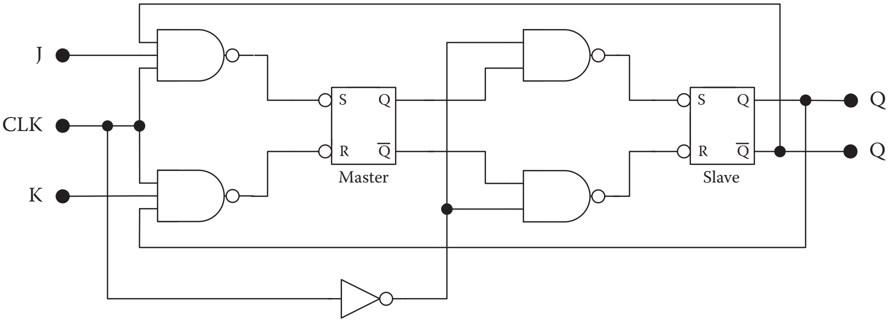


FIGURE 10.14
Master-slave JK flip-flop.

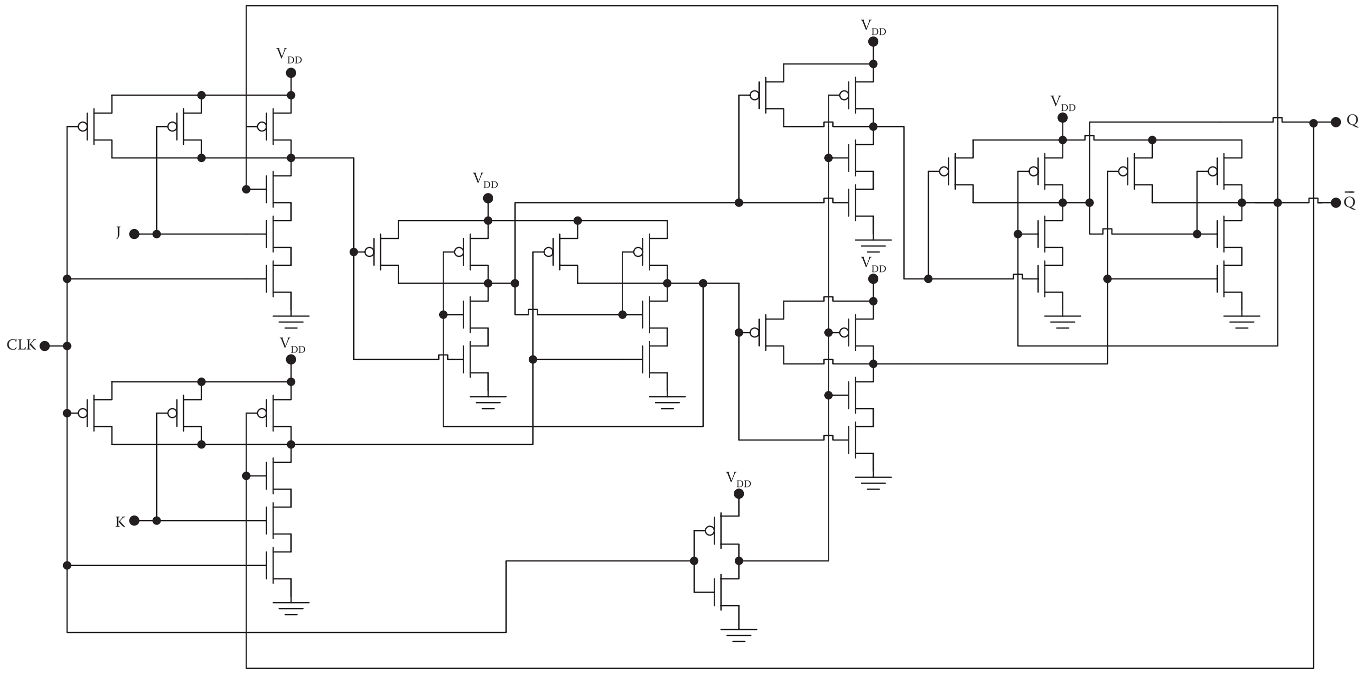


FIGURE 10.15
CMOS master-slave JK flip-flop.

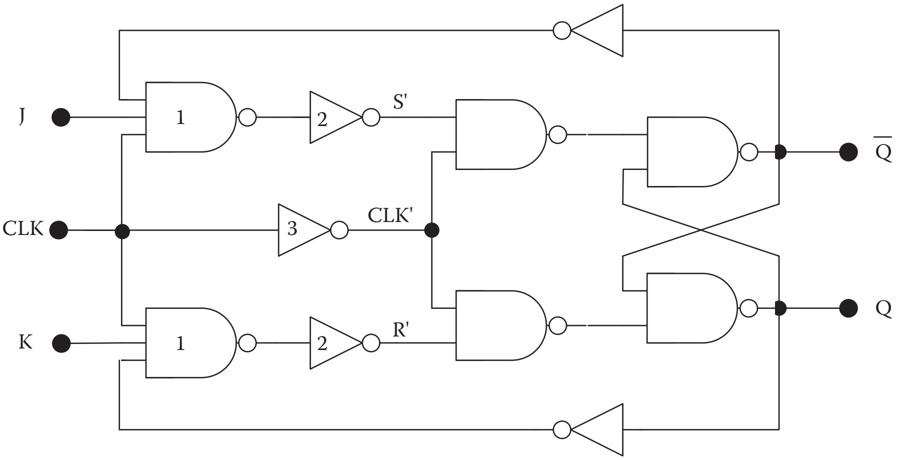


FIGURE 10.16
Edge-triggered JK flip-flop.

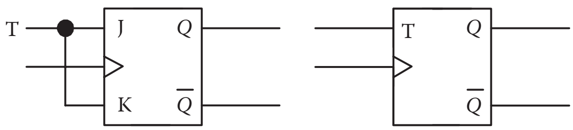
Although the master-slave and edge-triggering concepts were explained separately for clarity, it is possible to incorporate both features in a single flip-flop. This is nearly always done in practice.

10.5 Other Flip-Flops

Several special purpose flip-flops complement the JK type devices. Some of these can be realized by specially connecting the JK flip-flops. Two examples are the T ("toggle") flip-flop and the D ("data") flip-flop.

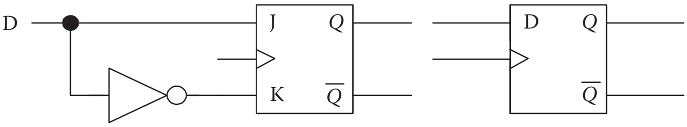
The T flip-flop may be realized by connecting together the J and K inputs of a JK flip-flop as shown in Figure 10.17. The application of logic one to the T input causes the output to toggle (change state) during each clock period. On the other hand, the application of logic zero to the T input causes 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 flip-flop as shown in Figure 10.18. 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 1 to the D input is characterized by J = 1 and K = 0. This causes the next state of the output to be 1. However, the application of logic 0 to the D input is characterized by J = 0 and K = 1. This causes 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.



T	Q_{N+1}
0	Q_N
1	$\overline{Q_N}$

FIGURE 10.17
T flip-flop.



D	Q_{N+1}
0	0
1	1

FIGURE 10.18
D flip-flop.

As with JK flip-flops, T and D flip-flops may use the master-slave and edge-triggered concepts.

10.6 Schmitt Triggers

Schmitt triggers are specially constructed bistable circuits that exhibit hysteresis; therefore, 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 sum of the noise margins may actually exceed the supply voltage.

In a Schmitt trigger, positive feedback and greater-than-unity loop gain are required, as with any bistable circuits. The achievement of hysteresis also requires that there be a switching element that introduces a state-dependent voltage between the input and ground.

The hysteresis characteristic of a general Schmitt trigger inverter is shown in Figure 10.19. 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 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 just *the hysteresis*:

$$V_H = V_U - V_L. \tag{10.3}$$

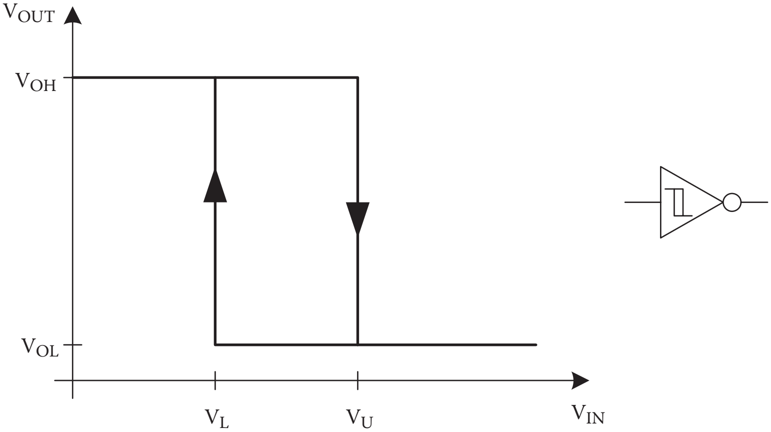


FIGURE 10.19
Schmitt trigger voltage transfer characteristic.

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 10.19.

In a static logic gate exhibiting no hysteresis, the noise margins are given by

$$V_{NMH} = V_{OH} - V_{IH} \tag{10.4}$$

and

$$V_{NML} = V_{IL} - V_{OL} \,. \tag{10.5}$$

It is always true that $V_{IL} \leq V_{IH}$, $V_{OH} \leq V_{DD}$, and $V_{IL} \geq 0$. Therefore, for a non-hysteresis gate, the sum of the noise margins can be no greater than the supply voltage:

$$V_{NML} + V_{NMH} \leq V_{DD} \,. \tag{10.6}$$

For a Schmitt trigger, this restriction on the sum of the noise margins is lifted. The modified noise margins for a Schmitt trigger are

$$V_{NMH} = V_{OH} - V_L \tag{10.7}$$

and

$$V_{NML} = V_U - V_{OL} \,. \tag{10.8}$$

Therefore, the sum of the noise margins for the Schmitt trigger is

$$V_{NML} + V_{NMH} = (V_{OH} - V_{OL}) + (V_U - V_L)\,. \tag{10.9}$$

For a Schmitt trigger exhibiting rail-to-rail logic swing,

$$V_{NML} + V_{NMH} = V_{DD} + V_H \text{ (Schmitt trigger),} \tag{10.10}$$

and because the hysteresis voltage can be as high as V_{DD},

$$V_{NML} + V_{NMH} \le 2V_{DD} \text{ (Schmitt trigger).} \tag{10.11}$$

The noise rejection afforded by hysteresis is especially important if a signal is to be applied to a count-up or count-down circuit. This can be understood by considering what happens when a noisy, slowly varying signal is applied to both a conventional inverter and a Schmitt trigger inverter as shown in Figure 10.20. The Schmitt trigger correctly interprets the waveform as a single

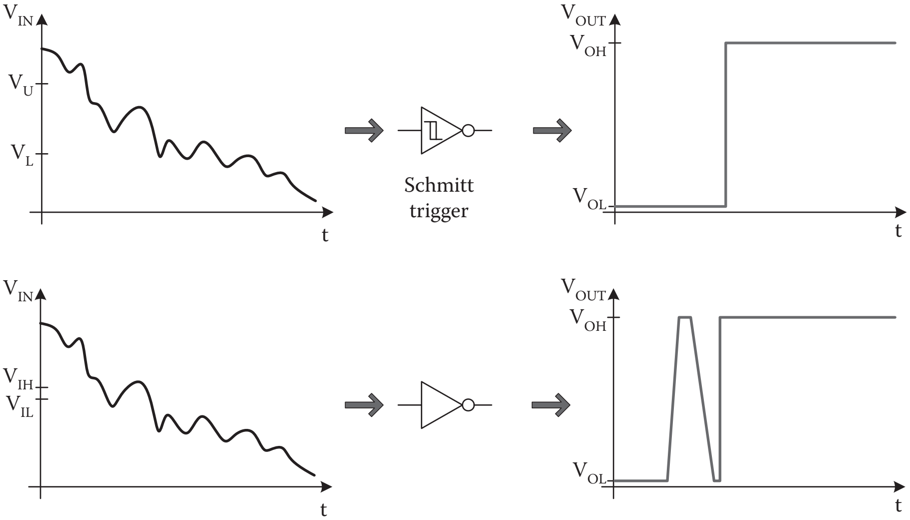


FIGURE 10.20
A noisy, slowly varying signal is applied to a Schmitt trigger (top) and a conventional inverter (bottom).

high-to-low transition, whereas the non-hysteresis inverter misinterprets the input waveform. Clearly, 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 increased short-circuit conduction and the associated dissipation.

10.6.1 CMOS Schmitt Trigger

Many different hysteresis circuits have been developed, but the most common CMOS realization is the six-transistor circuit shown in Figure 10.21.

To determine the voltage transfer characteristic for the CMOS Schmitt trigger, we will assume that all n-MOS transistors have a (positive) threshold voltage of V_{TN} and all p-MOS transistors have a (negative) threshold voltage of V_{TP}. The device transconductance value for M_{NI} is K_{NI}; for M_{NF}, it is K_{NF},

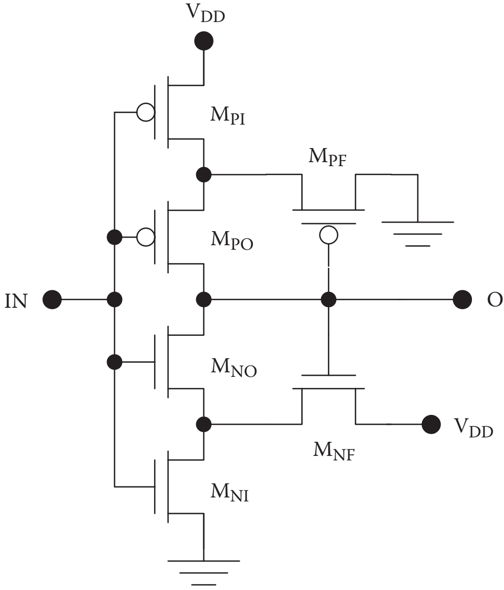


FIGURE 10.21
CMOS Schmitt trigger.

and so on. With $V_{IN} = 0$, M$_{NO}$, M$_{NI}$ and the feedback transistors M$_{PF}$ and M$_{NF}$ are cutoff, whereas M$_{PO}$ and M$_{PI}$ are linear. Therefore,

$$V_{OH} = V_{DD}\,. \tag{10.12}$$

If V$_{IN}$ is increased above V$_{TN}$, M$_{NI}$ and M$_{NF}$ will both be saturated. Together, these devices act like an NMOS inverter with a saturated enhancement type pull-up transistor. As long as the transistor M$_{NO}$ does not conduct, we can equate the drain currents of M$_{NI}$ and M$_{NF}$:

$$\frac{K_{NI}}{2}\left(V_{IN} - V_{TN}\right)^2 = \frac{K_{NF}}{2}\left(V_{GSNF} - V_{TN}\right)^2\,. \tag{10.13}$$

Solving, the gate to source voltage for the n-channel feedback transistor is

$$V_{GSNF} = \sqrt{\frac{K_{NI}}{K_{NF}}}\left(V_{IN} - V_{TN}\right) + V_{TN}\,. \tag{10.14}$$

The drain-to-source voltage for M$_{NI}$ is therefore

$$V_{DSNI} = V_{DD} - V_{GSNF} = V_{DD} - \sqrt{\frac{K_{NI}}{K_{NF}}}\left(V_{IN} - V_{TN}\right) - V_{TN}\,. \tag{10.15}$$

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,

$$V_{GSNO} = V_{IN} - V_{DSNI} = V_{TN}\,. \tag{10.16}$$

Solving, we find the upper trip voltage to be

$$V_U = \frac{V_{DD} + V_{TN}\sqrt{\dfrac{K_{NI}}{K_{NF}}}}{1 + \sqrt{\dfrac{K_{NI}}{K_{NF}}}}\,. \tag{10.17}$$

For the determination of the lower trip voltage, suppose that the input voltage is decreased starting from V$_{DD}$. With $V_{IN} = V_{DD}$, M$_{PO}$, M$_{PI}$ and the feedback transistors M$_{PF}$ and M$_{NF}$ are cutoff, whereas M$_{NO}$ and M$_{NI}$ are linear. Therefore,

$$V_{OL} = 0 \cdot \tag{10.18}$$

If V_{IN} is decreased below $V_{DD} - V_T$, M_{PI} and M_{PF} will both 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, we can equate the drain currents of M_{PI} and M_{PF}:

$$\frac{K_{PI}}{2}\left(V_{DD} - V_{IN} + V_{TP}\right)^2 = \frac{K_{PF}}{2}\left(V_{GSPF} - V_{TP}\right)^2 . \tag{10.19}$$

Solving, the gate-to-source voltage for the p-channel feedback transistor is given by

$$-V_{GSPF} = \sqrt{\frac{K_{PI}}{K_{PF}}}\left(V_{DD} - V_{IN} + V_{TP}\right) - V_{TP} . \tag{10.20}$$

Therefore, the voltage at the source of M_{PO} with respect to ground is also

$$V_{SPO} = \sqrt{\frac{K_{PI}}{K_{PF}}}\left(V_{DD} - V_{IN} + V_{TP}\right) - V_{TP} . \tag{10.21}$$

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,

$$V_{GSPO} = V_{IN} - V_{SPO} = V_{TP} . \tag{10.22}$$

Solving, we find the lower trip voltage to be

$$V_L = \frac{\left(V_{DD} + V_{TP}\right)\sqrt{\dfrac{K_{PI}}{K_{PF}}}}{1 + \sqrt{\dfrac{K_{PI}}{K_{PF}}}} . \tag{10.23}$$

* *PMOS* is a logic family that preceded NMOS. The circuits are built using only p-MOS transistors. The inverter uses 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-MOS transistors, and even PMOS microprocessors appeared on the market. However, as soon as the technology permitted, PMOS was displaced by superior NMOS circuitry.

Example 10.2 CMOS Schmitt Trigger

Determine the voltage transfer characteristic for the CMOS Schmitt trigger of Figure 10.22.

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*)

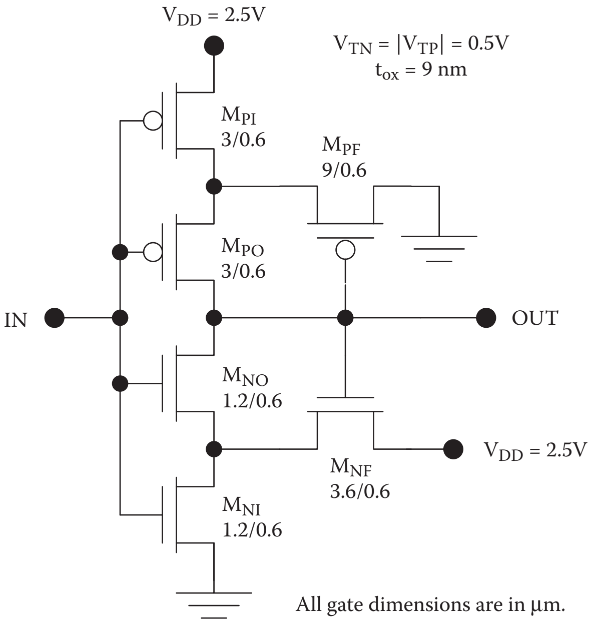


FIGURE 10.22
Example CMOS Schmitt trigger.

The output voltage levels are

$$V_{OL} = 0$$

and

$$V_{OH} = 2.5V.$$

The trip voltages are

$$V_U = \frac{V_{DD} + V_{TN}\sqrt{\dfrac{K_{NI}}{K_{NF}}}}{1+\sqrt{\dfrac{K_{NI}}{K_{NF}}}} = \frac{2.5V + 0.5V\sqrt{1/3}}{1+\sqrt{1/3}} = 1.77V$$

and

$$V_L = \frac{\left(V_{DD} + V_{TP}\right)\sqrt{\dfrac{K_{PI}}{K_{PF}}}}{1+\sqrt{\dfrac{K_{PI}}{K_{PF}}}} = \frac{\left(2.5V - 0.5V\right)\sqrt{1/3}}{1+\sqrt{1/3}} = 0.73V \cdot$$

Therefore, the circuit exhibits hysteresis of 1.04 V. The noise margins are $V_{NML} = 1.77V - 0 = 1.77V$, and $V_{NMH} = 2.5V - 0.73V = 1.77Vw$, so V_{NML} + V_{NMH} = 3.54V; this is about 42% greater than V_{DD}. The voltage transfer characteristic is shown in Figure 10.23.

Example 10.3. Sizing of the Feedback Transistors in a CMOS Schmitt Trigger

For the CMOS Schmitt trigger of Figure 10.24, determine the upper trip voltage as a function of the W_{PF}/W_{PI} ratio and the lower trip voltage as a function of W_{NF}/W_{NI}.

Solution: The trip voltages are

$$V_U = \frac{V_{DD} + V_{TN}\sqrt{\dfrac{K_{NI}}{K_{NF}}}}{1+\sqrt{\dfrac{K_{NI}}{K_{NF}}}} = \frac{V_{DD}\sqrt{\dfrac{W_{NF}}{W_{NI}}} + V_{TN}}{\sqrt{\dfrac{W_{NF}}{W_{NI}}} + 1} = \frac{2.5V\sqrt{\dfrac{W_{NF}}{W_{NI}}} + 0.5V}{\sqrt{\dfrac{W_{NF}}{W_{NI}}} + 1}$$

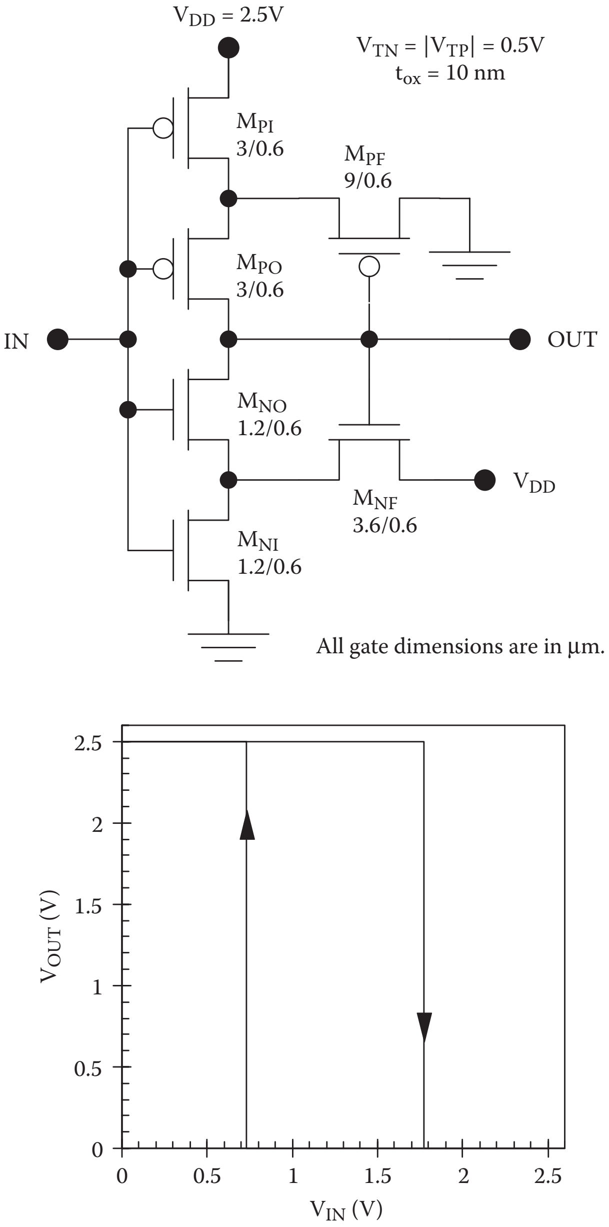


FIGURE 10.23
Example CMOS Schmitt trigger and voltage transfer characteristic.

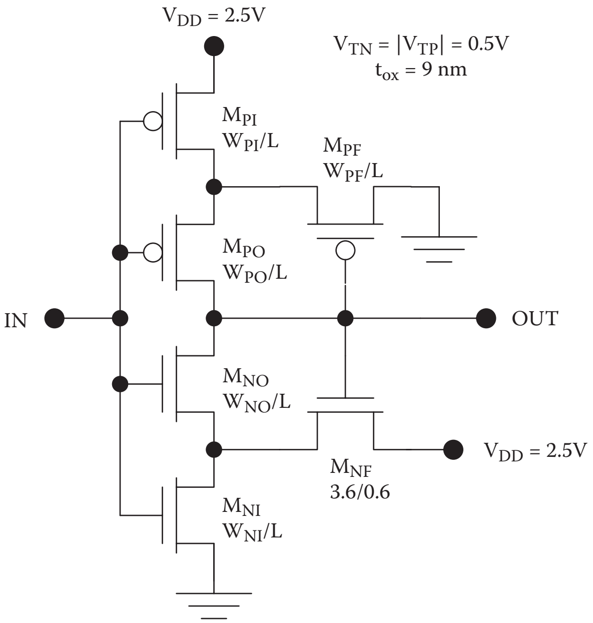


FIGURE 10.24
Example CMOS Schmitt trigger.

and

$$V_L = \frac{(V_{DD} + V_{TP})\sqrt{\dfrac{K_{PI}}{K_{PF}}}}{1 + \sqrt{\dfrac{K_{PI}}{K_{PF}}}} = \frac{V_{DD} + V_{TP}}{\sqrt{\dfrac{W_{PF}}{W_{PI}}} + 1} = \frac{2.0V}{\sqrt{\dfrac{W_{PF}}{W_{PI}}} + 1}.$$

The results plotted in Figure 10.25 show that the feedback transistors must be made wider than the other devices in the circuit to obtain a useful Schmitt trigger.

10.7 SPICE Demonstrations

For the purpose of illustration, simulations were performed using Cadence Capture CIS 10.1.0 PSpice (Cad ence Design Systems). The level 1 MOS transistor model parameters given in Tables 10.1 and 10.2 were used unless otherwise noted. The process transconductance parameters were calculated assuming an oxide thickness of 9 nm. For n-MOSFETS,

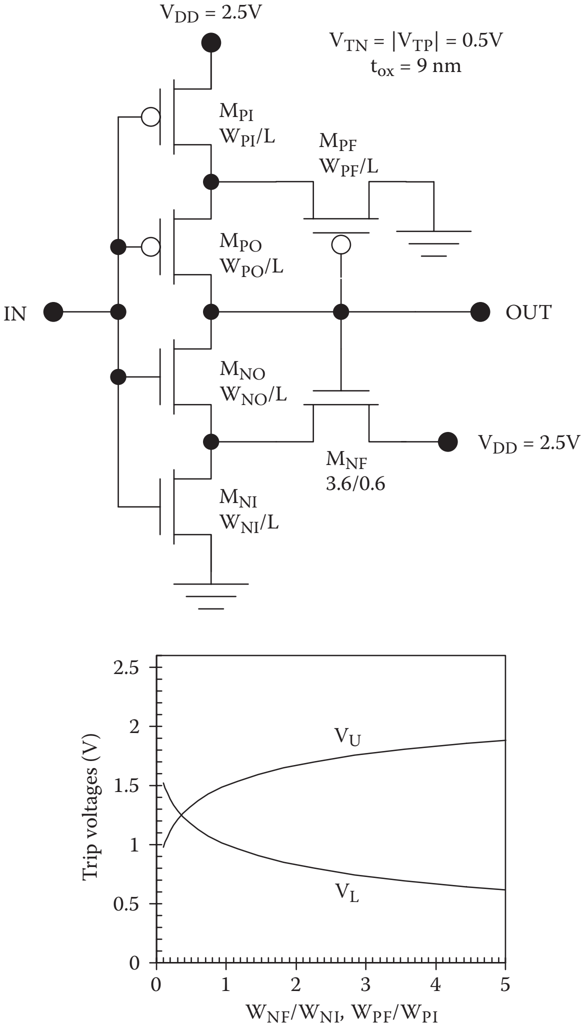


FIGURE 10.25
Calculated trip voltages for a CMOS Schmitt trigger as functions of WNF/WNI¬ and WPF/WPI.

$$KP = \frac{(3.9)\left(8.85\times10^{-14}F/cm\right)\left(580cm^2V^{-1}s^{-1}\right)}{9\times10^{-7}cm} = 222\mu A/V^2, \quad (10.24)$$

and for p-MOSFETS,

TABLE 10.1

n-MOS Level 1 SPICE Parameters

Parameter	Value	Units
KP	222u	A/V^2
VTO	0.5	V
GAMMA	0.15	V$^{1/2}$
PHI	0.7	V
LAMBDA	0.05	
TOX	9n	m
NSUB	1E16	cm^{-3}
UO	580	cm^2/Vs
CGSO	0.38n	F/m
CGDO	0.38n	F/m

$$KP = \frac{(3.9)\left(8.85\times10^{-14}F/cm\right)\left(230cm^2V^{-1}s^{-1}\right)}{9\times10^{-7}cm} = 88\mu A/V^2\,. \tag{10.25}$$

The overlap capacitances per unit gate width were determined with the assumption that $L_{OV} = 0.1\mu m$:

$$\begin{aligned} CGSO &= \frac{(3.9)\left(8.85\times10^{-14}F/cm\right)\left(0.1\times10^{-4}cm\right)}{9\times10^{-7}cm} \\ &= 3.8pF/cm = 0.38nF/m \end{aligned} \tag{10.26}$$

TABLE 10.2

p-MOS Level 1 SPICE Parameters

Parameter	Value	Units
KP	88u	A/V^2
VTO	–0.5	V
GAMMA	0.15	V$^{1/2}$
PHI	0.7	V
LAMBDA	0.05	
TOX	9n	m
NSUB	1E16	cm^{-3}
UO	580	cm^2/Vs
CGSO	0.38n	F/m
CGDO	0.38n	F/m

and

$$CGDO = \frac{(3.9)\left(8.85\times10^{-14}F/cm\right)\left(0.1\times10^{-4}cm\right)}{9\times10^{-7}cm}$$
$$= 3.8pF/cm = 0.38nF/m. \tag{10.27}$$

The body effect coefficient was calculated from

$$GAMMA = \frac{\sqrt{2q\varepsilon_{Si}N_a}}{C_{ox}}$$
$$= \frac{\sqrt{2\left(1.602\times10^{-19}C\right)(11.9)\left(8.85\times10^{-14}F/cm\right)\left(10^{16}cm^{-3}\right)}}{(3.9)\left(8.85\times10^{-14}F/cm\right)/9\times10^{-7}cm} \tag{10.28}$$
$$\approx 0.15V^{1/2}.$$

SPICE Example 10.1 CMOS Schmitt Trigger

Two DC sweeps, one positive going and one negative going, were used to determine the voltage transfer characteristics for the CMOS Schmitt trigger shown in Figure 10.26. In this circuit, the feedback transistors are three times as wide as the other transistors in the circuit. The composite characteristic of Figure 10.27 shows trip voltages of 0.7 and 1.8 V.

SPICE Example 10.2 CMOS Schmitt Triggers: Effect of V_{DD}

DC sweeps were used to determine the characteristics for CMOS Schmitt triggers of the design shown in Figure 10.28 using different supply voltages (1.5, 2.0, and 2.5 V). Figure 10.29 shows the trip voltages obtained from this analysis; both trip voltages depend on the supply voltage, but the upper trip voltage has a stronger dependence.

SPICE Example 10.3 CMOS Schmitt Triggers: Effect of Feedback Transistor Widths

DC sweeps were used to determine the characteristics for CMOS Schmitt triggers of the design shown in Figure 10.30 using three different values of the width ratio K (where $K = W_{PF}/W_{PI} = W_{NF}/W_{NI}$). Increasing the width ratio increases the hysteresis $V_U - V_L$ as shown in Figure 10.31.

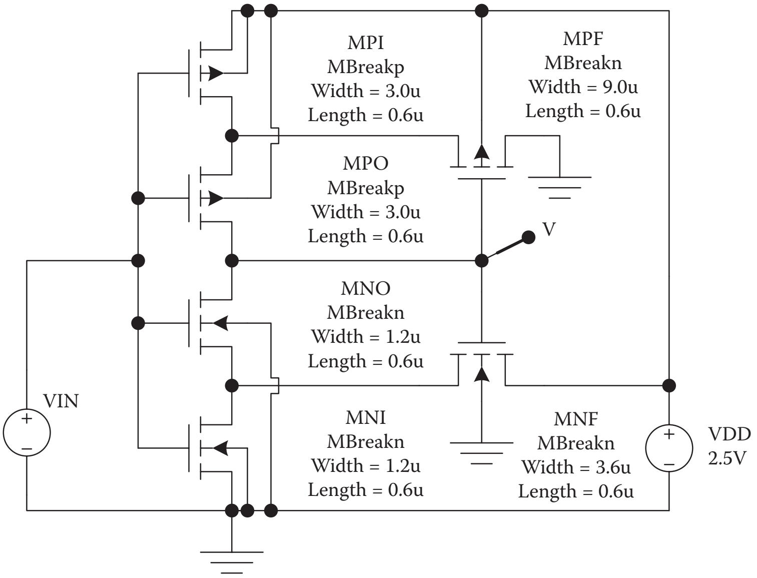


FIGURE 10.26
CMOS Schmitt trigger for the determination of the voltage transfer characteristics.

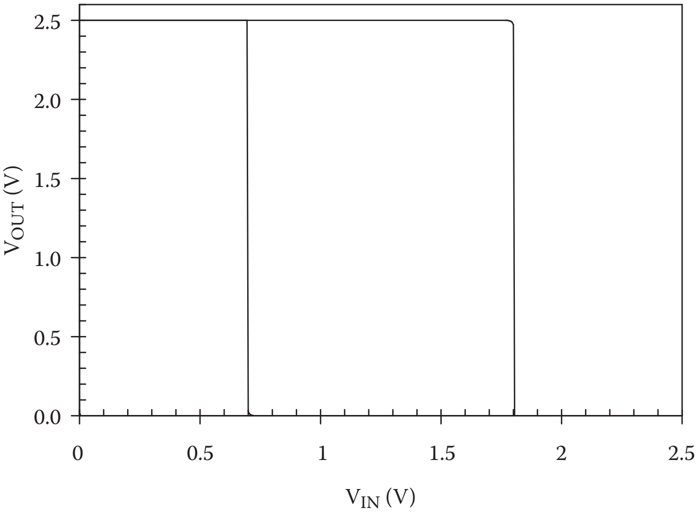


FIGURE 10.27
Composite voltage transfer characteristic for the CMOS Schmitt trigger of Figure 10.26, constructed by combining the results of two DC sweeps.

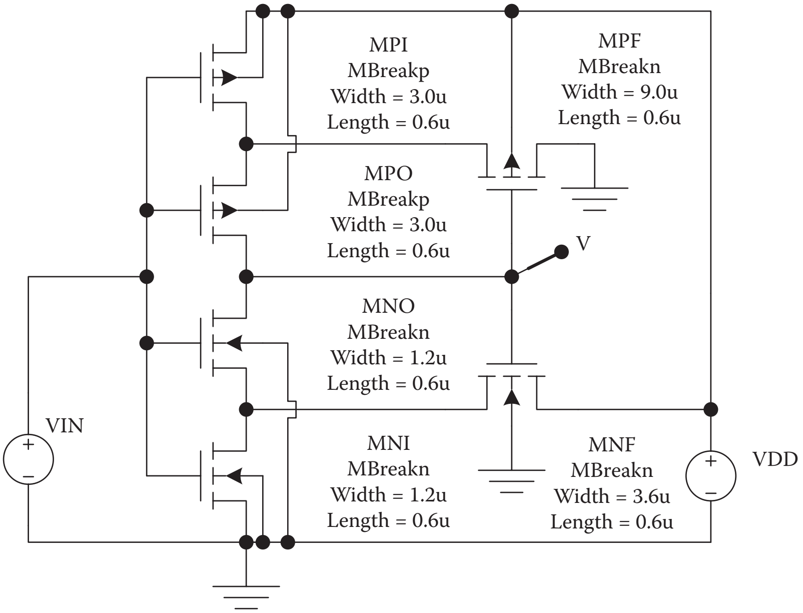


FIGURE 10.28
CMOS Schmitt trigger for the determination of the trip voltages for different values of the supply voltage.

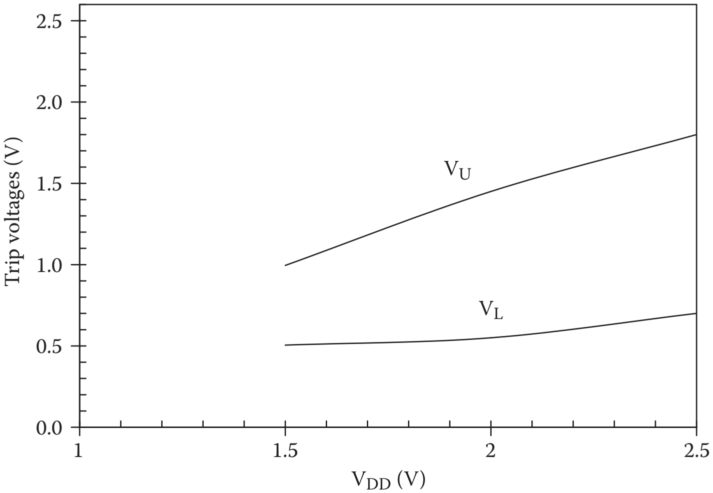


FIGURE 10.29
Trip voltages as functions of the supply voltage for CMOS Schmitt triggers with the design shown in Figure 10.28.

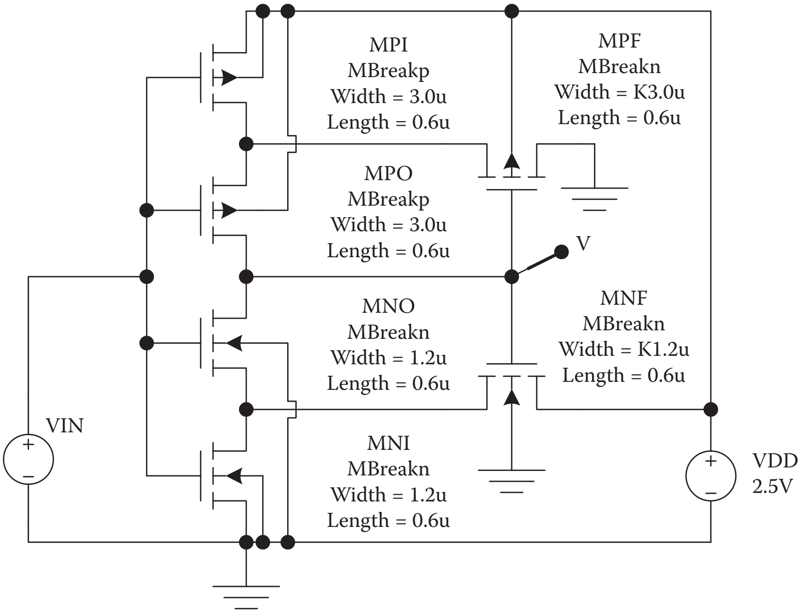


FIGURE 10.30
CMOS Schmitt trigger for the determination of the trip voltages with different values of the transistor width ratio K (where).

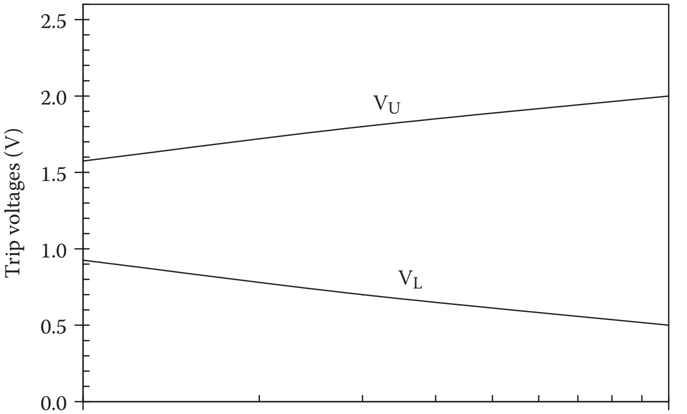


FIGURE 10.31
Trip voltages as functions of the transistor width ratio K (where) for CMOS Schmitt triggers of the design shown in Figure 10.30.

10.8 Practical Perspective

For practical perspective articles, see the dynamic website at http://www.engr.uconn.edu/ece/books/ayers.

10.9 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, and they are useful in signal-shaping applications.

10.10 Exercises

E10.1. Determine the trip voltage for the CMOS Schmitt trigger of Figure 10.32.

E10.2. Determine and plot the voltage transfer characteristics for the circuit of Figure 10.33, with $V_{DD} = 1.5$, 2, and 2.5 V.

E10.3. Choose widths for the MOS transistors in the circuit of Figure 10.34 so that the trip voltages are $V_U = 1.85V$ and $V_L = 0.9V$. $L = 0.6\mu m$

For additional exercise problems, see the dynamic website at http://www.engr.uconn.edu/ece/books/ayers.

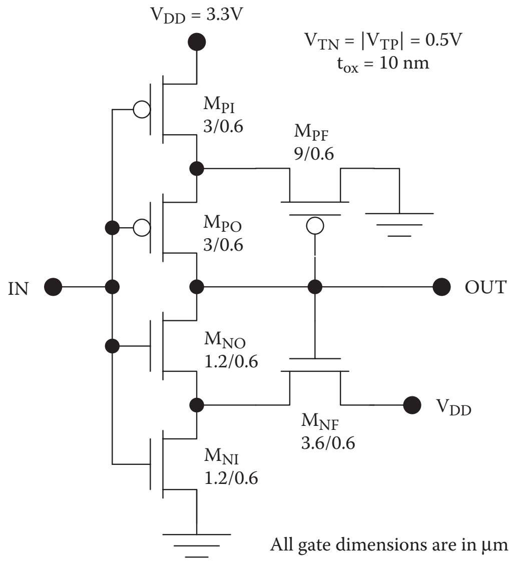


FIGURE 10.32
CMOS Schmitt trigger for the determination of the trip voltages (see Exercise E10.1).

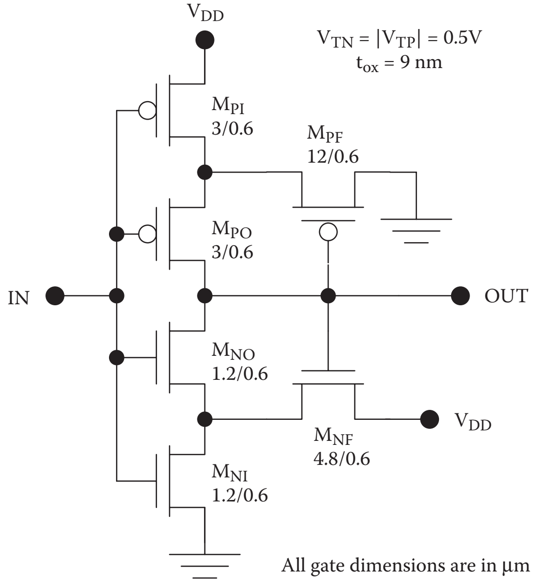


FIGURE 10.33
Schmitt trigger with varied supply voltage (see Exercise E10.2).

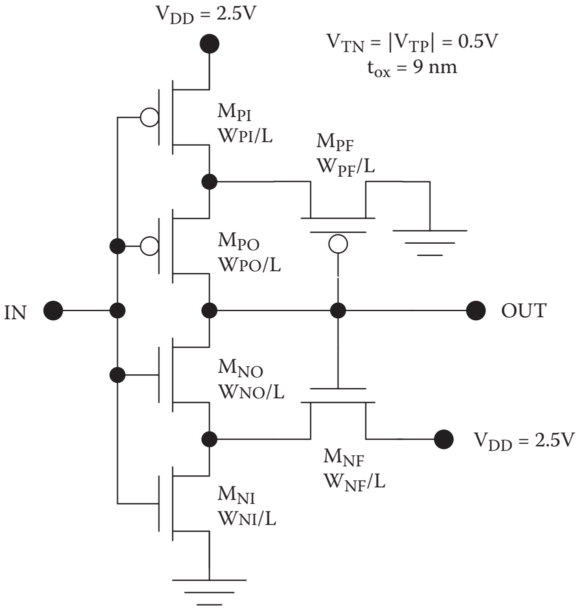


FIGURE 10.34
Schmitt trigger for the design of the transistor widths (see Exercise E10.3).

References

1. Baker, R. J., *CMOS circuit design, layout, and simulation,* 2nd ed., IEEE Press, Piscataway, NJ, 2007.
2. Yuan, J., and Svensson, C., New single-clock CMOS latches and flipflops with improved speed and power savings. *IEEE Solid-State Circuits*, 32, 62–69, 1997.
3. Filanovsky, I. M., and Baltes, H., CMOS Schmitt trigger design. *IEEE Trans. Circuits Syst.*, 41, 46–49, 1994.

11

Digital Memories

11.1 Introduction

Digital memories store bits of information for later use by processors, displays, and input/output devices and are therefore crucial elements in most digital systems. Broadly speaking, digital memories can be classified as volatile or nonvolatile. Nonvolatile memories retain their data when the system power is turned off and are necessary for storing documents, images, and video files. Nonvolatile memory is also used to store the startup system for a processor. Computers and servers make extensive use of volatile memory that is loaded with programs and files 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 a large digital system, the data storage is organized according to capacity and speed. A limited amount of high-performance memory is located on the processor chip itself, whereas highest capacity media such as magnetic and optical drives (which are also the slowest) are farthest from the processor. Intermediate types of storage include volatile and nonvolatile memory circuits that are considered in this chapter.

All digital memory circuits use the organization shown in Figure 11.1, with memory cells 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 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 jth row and the kth 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 interpreted by the column decoder; in addition, the column decoding circuitry is responsible for transferring data in and out of the memory. For this reason, column lines are also known as *bit lines*. Row lines, on the other hand, select entire rows (or words) of data so these are called *word lines*. The core of the memory may be split into a number of blocks, to keep the lengths of the word and bit lines manageable. This is important because the access time for a memory is often

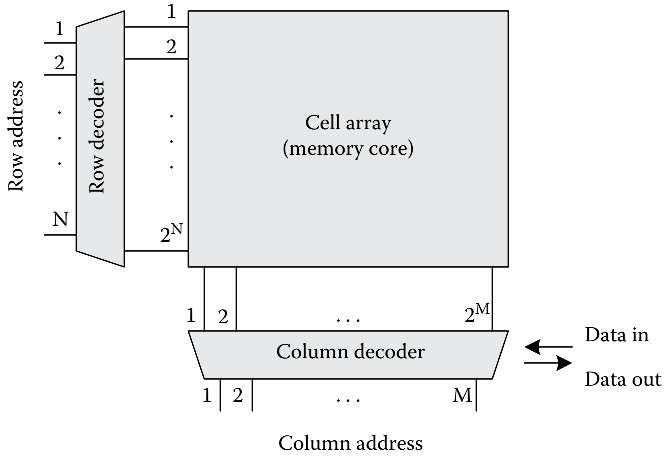


FIGURE 11.1
General design of a digital memory.

limited by the delays associated with the long interconnects serving as bit and word lines.

Digital memories are the integrated circuits with the highest level of integration. This is a consequence of two important factors. First, the core of a flash memory or DRAM contains many simple and identical cells, usually with one transistor per bit. It is therefore common to design in some 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 made the memory business (along with the microprocessor and application-specific integrated circuit industries) a technology driver for the entire silicon integrated circuit industry. Additional increases in memory density will come about by a combination of device scaling, increased die size, and the ability to store multiple bits with a single transistor.

Memory chips are classified as ROM or random access memory (RAM). The latter type should really be called "read write memory," because it is the capability to write data that 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.

RAMs can be further classified as static RAM (SRAM) and DRAM. SRAMs store information in latches. Therefore, these chips retain their data as long as the system power is on, without the need for clocking or refreshing. DRAMs 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 the loss of data. RAMs are inherently volatile in nature.

ROMs 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 removal of 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 fast erasing of large blocks of data. All ROMs are nonvolatile.

Today there is an almost limitless selection of memory circuits in the marketplace. No attempt will be made to catalog them here. Instead, we will focus on the general principles underlying memory circuits.

11.2 Static Random Access Memory

SRAMs use bistable latch circuits to store bits of data. A basic SRAM cell to store one bit comprises two cross-coupled inverters in a latch arrangement and two access devices as shown in Figure 11.2. Here R is the row (word) line and C, $\overline{C}$ are complementary column (bit) lines. Writing a bit involves bringing the word line high, thus turning on the two access devices, and then driving the bit lines to force the cross-coupled latch to one of its two

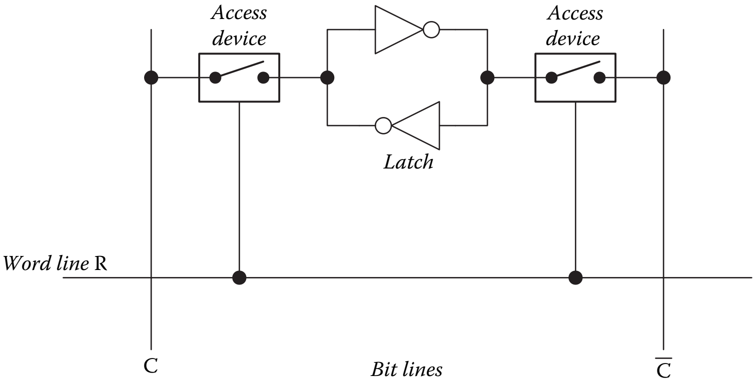


FIGURE 11.2
Basic SRAM cell comprising two cross-coupled inverters and two access devices.

stable states. To write a "1," C is forced to V_{DD}, whereas $\overline{C}$ is forced to 0. To write a "0," $C \to 0$ but $\overline{C} \to V_{DD}$. Once the writing operation has completed, the latch will remain in the forced state after the word line has gone low and the access devices have been turned off. Reading a bit involves bringing the word line high, to turn on the access devices, and then sensing either or both voltages on the bit lines. The adoption of two complementary bit lines enables use of a differential amplifier for this purpose.

11.2.1 CMOS SRAM Cell

In CMOS circuitry, SRAM cells are usually realized with cross-coupled CMOS inverters and n-MOS pass transistors as shown in Figure 11.3. This results in a cell with six transistors, which is referred to as a *6T SRAM cell* [1–3]. To write a "1," the word line is brought high, C is forced to V_{DD}, and $\overline{C}$ is forced to 0. This puts the latch in state 1, with M_{PO1} and M_{NO2} linear but M_{NO1} and M_{PO2} cutoff. When the word line is brought high in a subsequent read operation, the voltages sensed on the bit lines will be $C \approx V_{DD}$ and $\overline{C} \approx 0$. In fact, any small positive excursion of the voltage difference $V_C - V_{\overline{C}}$ can be interpreted as a "1" so that a high-gain amplifier can read the bit in a fraction of the bit line rise time. To write a "0," the word line is brought high, C is forced to 0, and

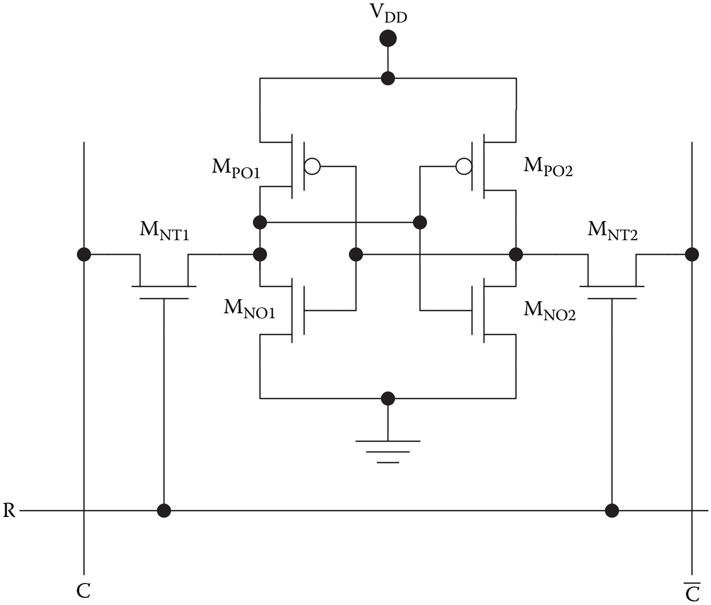


FIGURE 11.3
CMOS 6T SRAM cell.

$\overline{C}$ is forced to V_{DD}. This puts the latch in state 0, for which M_{PO1} and M_{NO2} are cutoff but M_{NO1} and M_{PO2} are linear. In a subsequent read operation, $V_C - V_{\overline{C}}$ will make a negative excursion when the word line is brought high.

The CMOS 6T SRAM cell uses cross-coupled inverters instead of a NAND- or NOR-type latch for the sake of a more compact layout. A NAND or NOR latch by itself would require eight transistors, resulting in a 10T cell that would require more area. A drawback of the simpler four-transistor latch is the need for increased bit line current drive whenever a write operation changes the latch's state. For example, suppose the latch is in state 0 (M_{PO1} and M_{NO2} are cutoff but M_{NO1} and M_{PO2} are linear). To write a "1," it is necessary to drive sufficient drain current in M_{NO1} so that its drain-to-source voltage reaches the threshold voltage for M_{NO2}. At the same time, the $\overline{C}$ line must sink enough current so that the drain-to-source voltage drop across M_{PO2} will be sufficient to that the gate voltage on M_{NO1} will drop below its threshold voltage.

The access devices in the CMOS 6T SRAM cell could be p-MOS pass transistors, with an active low word line, or CMOS transmission gates, driven by complementary word lines. However, these alternatives do not offer specific advantages in Si technology so more compact n-MOS pass transistors are used.

11.2.2 NMOS SRAM Cell

A 6T SRAM cell can also be realized using all n-MOS transistors as shown in Figure 11.4. The read/write operations for this memory cell are similar to those for the CMOS cell. A significant disadvantage of this cell is the static power dissipation because, regardless of the logic state, the cell draws a static supply current equal to $K_L V_{TL}^2 / 2$. The static dissipation in the CMOS cell is mostly attributable to subthreshold currents, which are orders of magnitude less.

11.2.3 SRAM Sense Amplifiers

The sense amplifiers used in CMOS SRAMs differ in design but typically CMOS differential amplifiers are used [4]. An important advantage of differential operation is the rejection of common-mode noise and interference, such as that caused by signal crosstalk or Ldi/dt induced voltages on the power connections. The ability of a differential amplifier to read a small voltage difference in the presence of noise means that the voltage can be read after a small fraction of the bit line rise time. Therefore, a high-gain sense amplifier allows shorter read times.

An example of a CMOS sense amplifier is shown in Figure 11.5. When the clock signal is low, M_{NCLK} is cutoff and so OUT and $\overline{OUT}$ precharge to $V_{DD} - |V_{TP}|$. 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 $\overline{C}$,

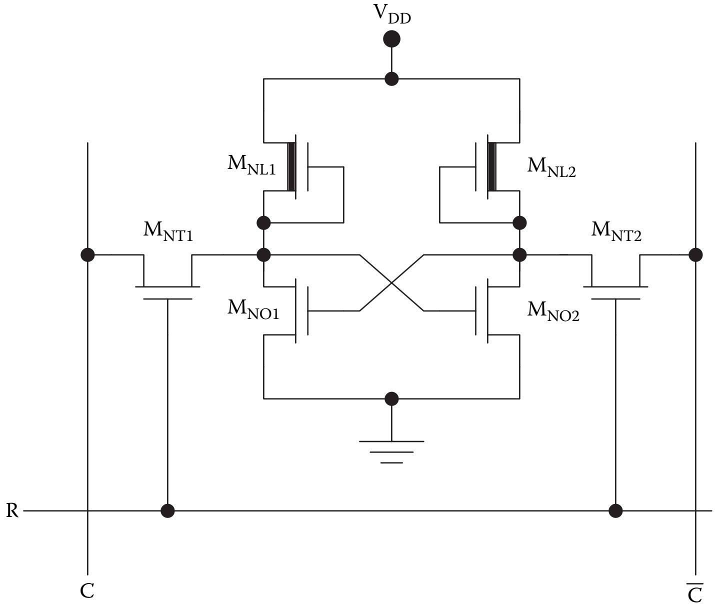


FIGURE 11.4
NMOS 6T SRAM cell.

M_{ND1} will turn on, M_{ND2} will turn off, and OUT 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 OUT will go low.

Because the read time is determined in great part by the voltage gain of the differential amplifier, it may be desirable to cascade two or more stages amplification. The availability of complementary outputs OUT and $\overline{OUT}$ facilitates this approach.

11.3 Dynamic Random Access Memory

DRAMs [5–7] achieve higher density than SRAMs 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 as shown in Figure 11.6. The bit stored in this cell is represented by the voltage on the storage capacitor C_S. (The right-hand capacitor in the figure is the column capacitance and is not associated with any one cell.) If the voltage on the storage capacitor is V_{DD}, a "1" is stored. On the other hand, logic "0" is represented by 0 V. At the present time, dedicated DRAM chips use 1T1C cells

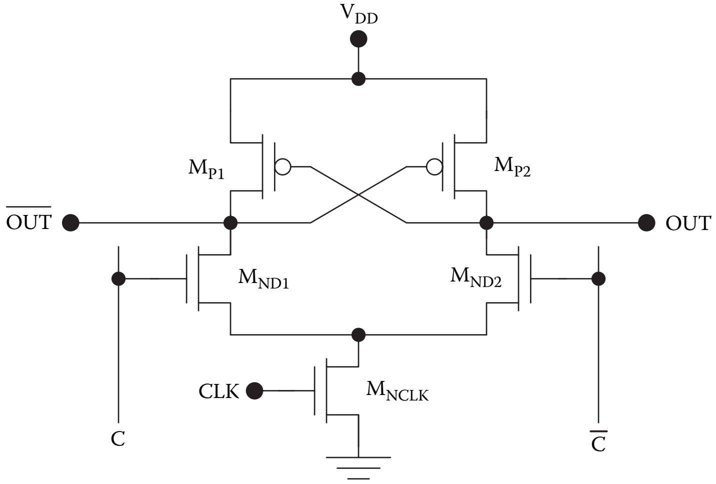


FIGURE 11.5
CMOS SRAM sense amplifier.

exclusively because of their high density. However, DRAMs that are embedded in systems on a chip often use two-transistor cells. These cells are less dense but provide improved signal-to-noise ratios and therefore greater reliability in the presence of crosstalk.

The operation of the 1T1C cell of Figure 11.6 is illustrated with the waveforms of Figure 11.7. During a write operation, the word line is bought high to turn on the access transistor and the column line is forced to either V_{DD} (for logic "1") or 0 (for logic "0") to set the voltage on the storage capacitor. During

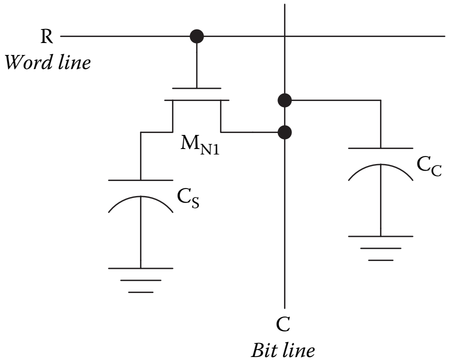


FIGURE 11.6
1T1C DRAM cell.

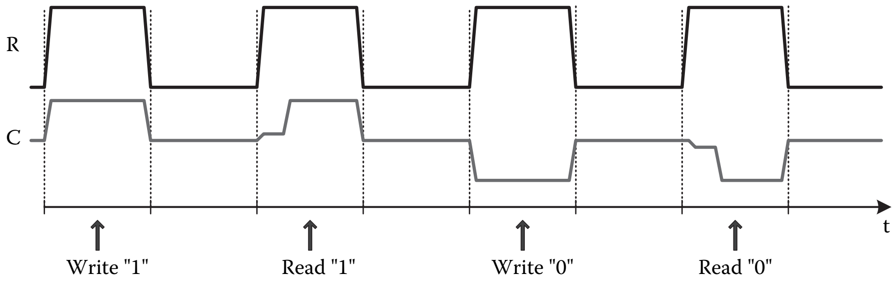


FIGURE 11.7
DRAM write and read waveforms.

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 "1" is inferred. If the column voltage drops below $V_{DD}/2$, then a "0" 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 sense amplifiers 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 0 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 complete.

The data stored in 1T1C DRAM cells must be refreshed periodically. This is attributable to the small leakage currents in the storage capacitors and access transistors. Typically, the leakage is of such a magnitude that the data must be refreshed every few milliseconds. Therefore, the column circuitry must perform read and refresh operations at a minimum frequency (~100 Hz) even when the stored data are not being fetched by the processor.

There are numerous designs for 1T1C DRAM read/refresh circuits. Generally, they all use differential amplifiers for the rejection of common mode signals as well as power supply glitches. However, because the 1T1C DRAM cell has a single-ended output, a special technique involving dummy cells is used [8–11].

Figure 11.8 illustrates this concept for the case of an NMOS differential amplifier, although any other differential amplifier could be used in much the same way. The column is split into two half-columns C and $\overline{C}$, and a reference cell (*dummy cell*) is associated with each half-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/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

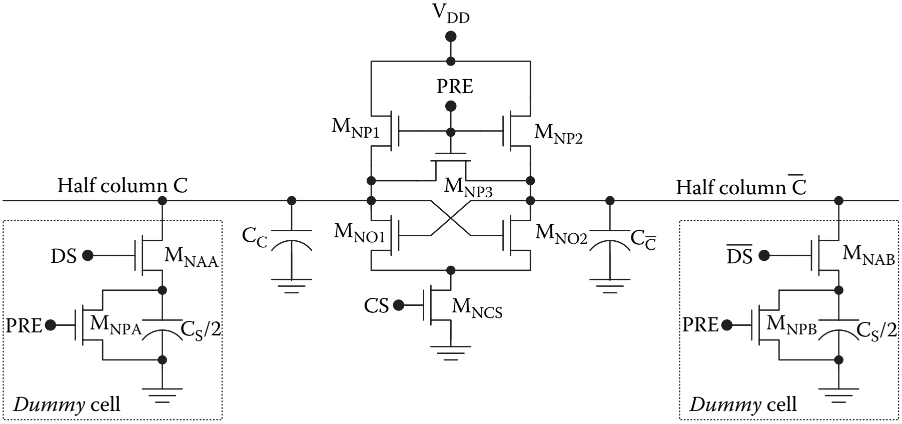


FIGURE 11. 8
DRAM read/refresh circuit.

selected is in half-column C, then $\overline{DS}$ is brought high.) Finally, the column is selected by bringing CS high. Then, 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 is read and that cell stored "1," 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 ground (GND). If a cell in half-column C is read and that cell stored "0," 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 the use of multiple arms or cylinders. Remarkably, it has proven possible to fabricate such structures with excellent levels of circuit yield.

Relentless device scaling will continue to increase the densities of DRAM chips for some time to come. Additional increases will also come about with steady increases in die sizes, made possible by the reduction in process defects. There is also considerable interest in eliminating the capacitor, resulting in a true 1T cell [12, 13]. Such a cell requires the storage of charge in the MOSFET structure itself and imposes considerable challenges. However, the 1T cell offers the potential for additional increases in the capacity of DRAM chips.

11.4 Read-Only Memory

ROM [14] is programmed at the factory and cannot be written to or reprogrammed once installed. Hence, 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 n-MOS transistor per bit, resulting in high density and low-power operation. Typically, the pattern of ones and zeros is programmed in a single masking step, to reduce cost and lead time. The basic configurations for ROMs can be classified as NOR or NAND designs.

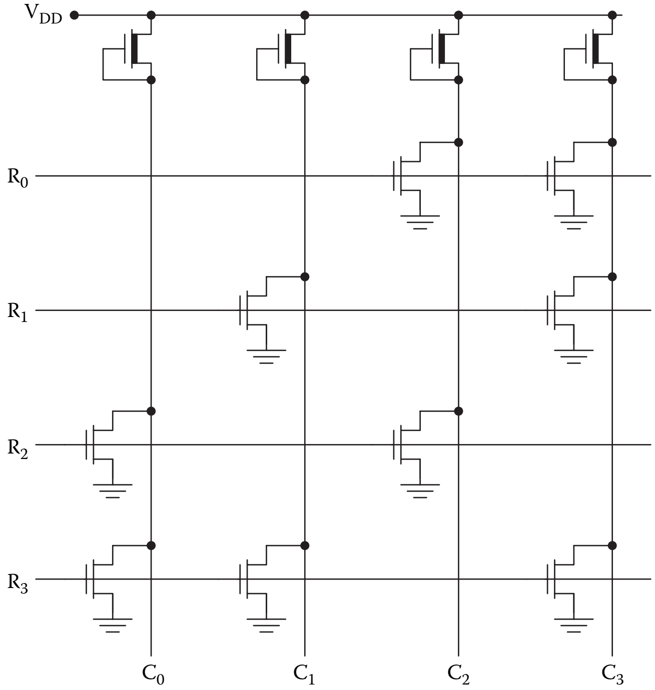
11.4.1 NOR Read-Only Memory

Figure 11.9 illustrates a 4×4 NMOS NOR ROM. Each bit line (column) has a depletion-type n-MOS pull-up transistor. Enhancement type n-MOS pull-down transistors are connected at cross points to store logic "0." If row R_0 is selected (brought high), the voltages at columns C_0 and C_1 will stay at V_{DD} because of the absence of pull-down transistors at the R_0C_0 and R_0C_1 cross points. Only columns C_2 and C_3 will go low (provided that only one row is selected at a time). The name of this type of circuit comes from the fact that each column acts as a NOR gate. In the example 4×4 memory shown in Figure 11.9,

$$\begin{aligned} C_0 &= \overline{R_2 + R_3} \\ C_1 &= \overline{R_1 + R_3} \\ C_2 &= \overline{R_0 + R_2} \\ C_3 &= \overline{R_0 + R_1 + R_3}. \end{aligned} \tag{11.1}$$

In a practical ROM, transistors are designed and fabricated at each cross point. The unwanted transistors are disabled in a single mask step that minimizes the extent of customization and therefore cost. For example, drain contacts from the metal wires forming the bit (column) lines may be omitted at the n-MOS transistors that are to be disabled. Because this step comes late in the fabrication process, wafers may be partially fabricated in anticipation of customer orders, with a reduction in the lead time. Figure 11.10 shows the layout for the NMOS NOR ROM of Figure 11.9. The polysilicon word lines form the gates of the cell transistors. The metal bit lines are connected to the drains of transistors only where zeros are to be stored. The ion-implanted ground regions may be shared by two adjacent rows. The pull-up transistors at the top have their gates connected to their sources.

A CMOS NOR ROM can be made using p-MOS pull-up devices as shown in Figure 11.11.



	C_0	C_1	C_2	C_3
R_0	1	1	0	0
R_1	1	0	1	0
R_2	0	1	0	1
R_3	0	0	1	0

FIGURE 11.9
4 × 4 NMOS NOR ROM.

11.4.2 NAND Read-Only Memory

The NAND ROM also requires one MOS device per bit, but the transistors within a column are arranged in series rather than parallel so that each column behaves as a NAND gate. Figure 11.12 illustrates a 4×4 NMOS NAND ROM. In contrast to the NOR design, the placement of a transistor represents logic "1," the word (row) lines are *active low*, and the column output voltages must be taken from the sources of the pull-up transistors.

The basic operation of the NMOS NAND ROM is as follows. The row voltages are normally 0. To select row one, $\overline{R_1}$ is brought low (0 V). This will turn off the transistors at the R_1C_1 and R_1C_2 cross points, so the voltages on

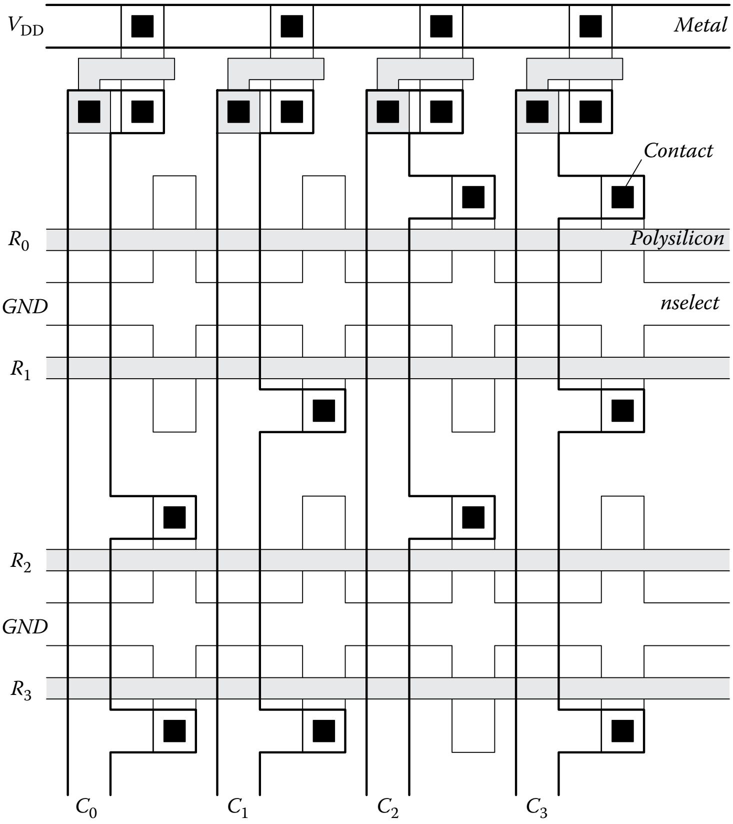


FIGURE 11.10
Layout design for a 4 × 4 NMOS NOR ROM.

columns C_1 and C_2 will go high (V_{DD}). All other column voltages will remain low (~0 V).

Here, too, transistors are designed and fabricated at each cross point. The transistors to be "eliminated" from the operation of the circuit may either be shorted by a metal column line or made into depletion-type devices by an ion implantation step. The latter approach renders the transistor so that it is on (linear) even when the word line is selected (brought to ~0 V). Therefore, the presence of such a depletion-mode transistor has little effect on the circuit operation.

Figure 11.13 shows a layout design for a NMOS NAND ROM in which the unneeded transistors have been rendered "always on" by an extra n-type ion implantation step. The word (row) lines are formed by polysilicon wires that serve as the gates of the cell transistors. Bit (column) lines are formed by the nselect implantation and so metal contacts are not necessary along their length. This important difference makes the NAND ROM layout more dense

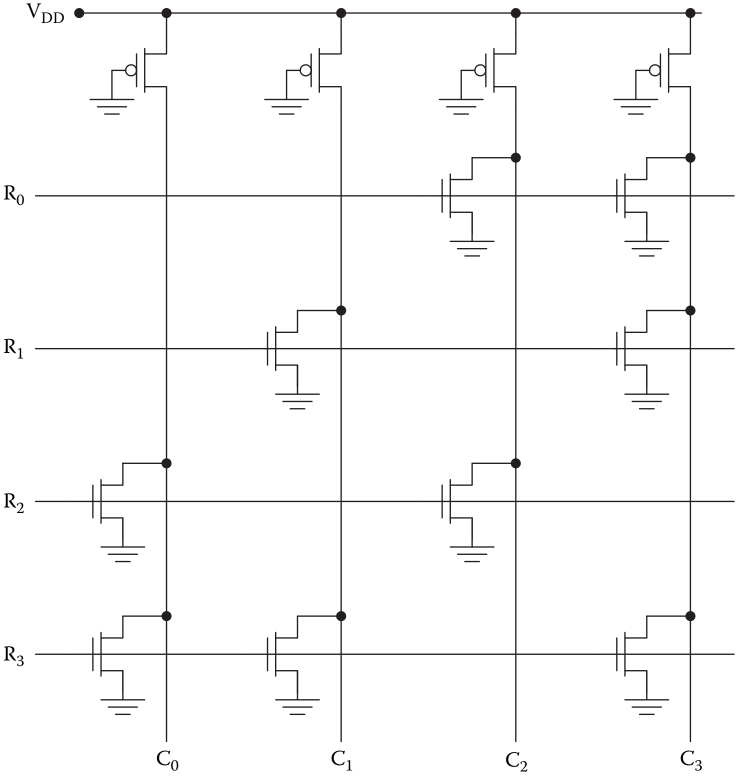


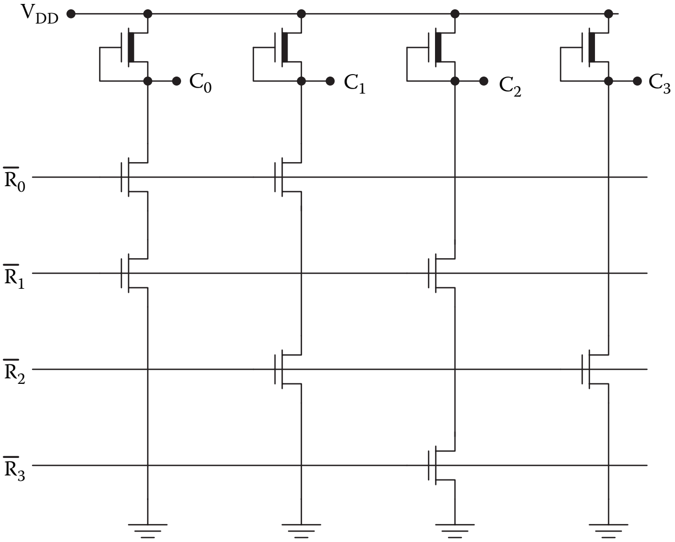
FIGURE 11.11
4 × 4 CMOS NOR ROM.

than the NOR ROM. However, the NAND ROM design is slower because of the series connected MOSFETs as well as the series resistance in the ion-implanted bit line.

A NAND ROM may be also be implemented using p-MOS pull up devices, and the operation of the circuit is essentially the same. Figure 11.14 illustrates such a CMOS NAND ROM.

11.5 Programmable Read-Only Memory

PROM was developed for the purposes of field-customized memories and prototyping. These circuits, using one transistor per bit, incorporated microfuses that could be blown selectively by the controlled application of electrical current. A drawback of PROM is that it could be programmed



	C_0	C_1	C_2	C_3
R_0	1	1	0	0
R_1	1	0	1	0
R_2	0	1	0	1
R_3	0	0	1	0

FIGURE 11.12
4 × 4 NMOS NAND ROM.

only once. These memories have now been rendered obsolete by the availability of EPROMs and flash memories, which are denser, faster, and more flexible than PROM.

11.6 Erasable Programmable Read-Only Memory

An EPROM allows multiple erase/program cycles, making it far more flexible than PROM. 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 11.15 [15]. As the name suggests, this device has a second gate that is floating (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

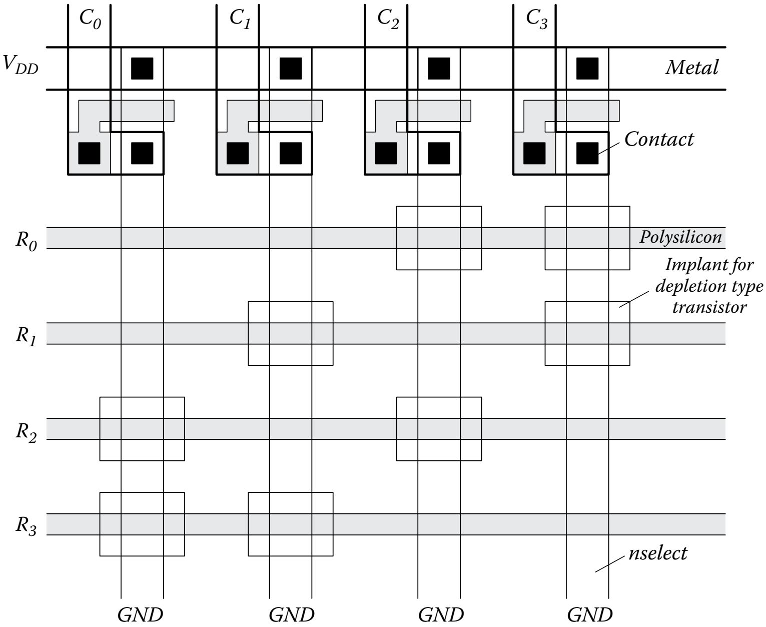


FIGURE 11.13
Layout design for a 4 × 4 NMOS NAND ROM.

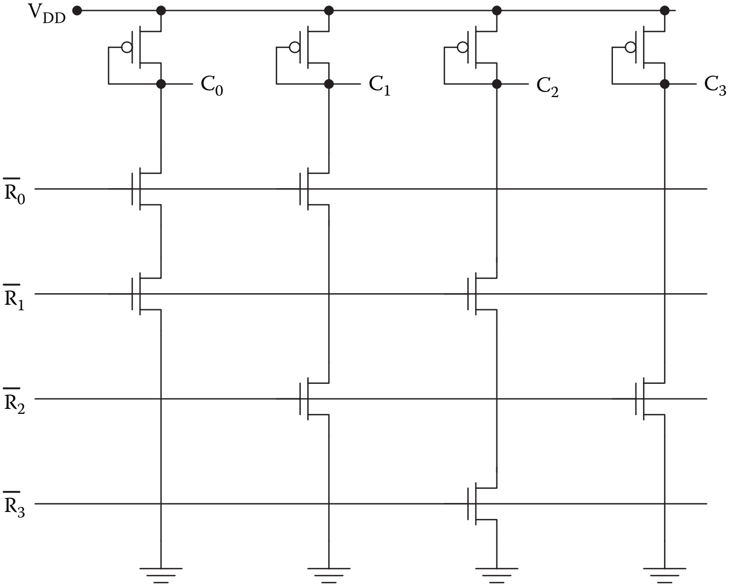


FIGURE 11.14
4 × 4 CMOS NAND ROM.

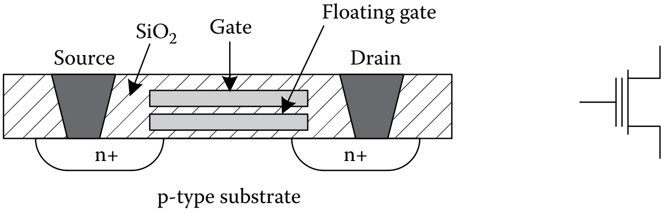


FIGURE 11.15
FAMOS.

gate repels electrons in the channel, making the threshold voltage more positive, whereas positive charge on the floating gate attracts electrons to the channel and makes the threshold voltage more negative. It is this property of the FAMOS that is exploited in EPROMs.

The circuit design for an EPROM is shown in Figure 11.16 and is essentially an NMOS NOR ROM constructed using floating gate MOSFETs. Logic "1" is programmed by making the threshold voltage of the appropriate FAMOS greater than V_{DD}. This ensures that the transistor will never turn on, and a FAMOS programmed in this way behaves as if it has been removed from the circuit. Logic "0" 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 "0." 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. 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 in which logic "1" must be stored. This is achieved by grounding the column line and placing large positive voltages on both 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* that they can be injected through the gate insulator to the floating gate. This is a self-limiting process,

* These high energy electrons are called "hot electrons" because they are not in thermal equilibrium with the semiconductor lattice. (Their average energy corresponds to that for electrons in equilibrium with a lattice at a temperature of 1000–10,000 K.)

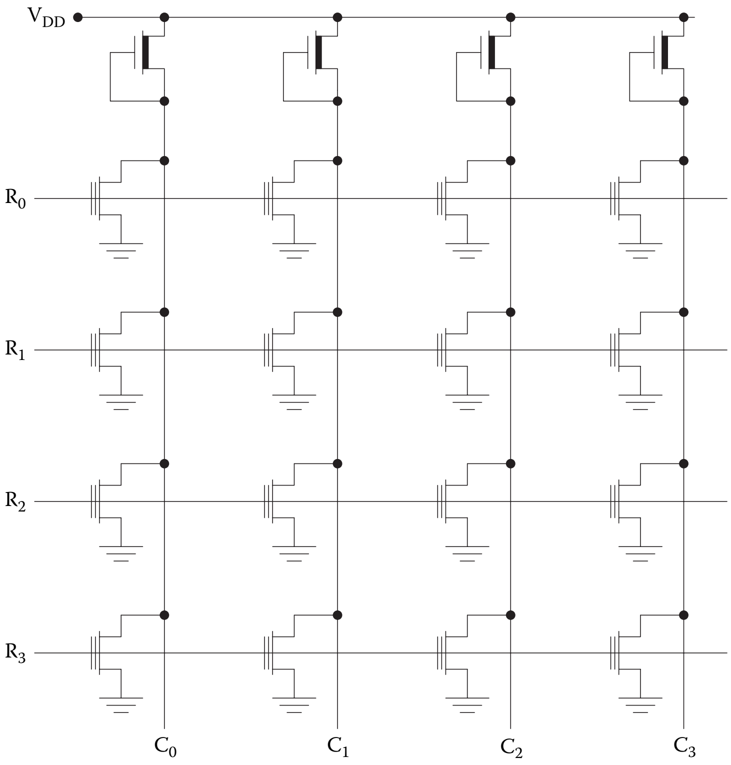


FIGURE 11.16
4 × 4 EPROM.

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 "1" at the selected location.

There are two drawbacks associated with EPROMs. The first is the inconvenience of erasure, which requires removal of the chip from the system to an ultraviolet flood chamber. The second is the inevitable oxide degradation that occurs with repeated erase/program cycles. This degradation, induced by the injection of hot electrons, gradually renders the oxide leaky and limits the total number of erase/program cycles to hundreds.

11.7 Electrically Erasable Programmable Read-Only Memory

The inconvenience of erasing EPROMs has led to the development of the EEPROMs (or E^2PROMs) [16]. The EEPROM is based on the floating gate tunnel oxide (FLOTOX) transistor illustrated in Figure 11.17. 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, there is a finite probability that the electron will spontaneously appear there. However, the wave function of an electron decays rapidly with distance and 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 the transport of electrons to and from the floating gate. During erasing/programming, electrons tunnel through this oxide in a number of intermediate jumps involving defect states in the oxide. Therefore, the specific process involved is called "Fowler–Nordheim tunneling."

The EEPROM circuit requires two transistors per bit as shown in Figure 11.18. Each cell contains one conventional MOSFET as well as a FLOTOX device. To program logic "1," sufficient negative charge is placed on the floating gate so that the threshold voltage ends up being greater than V_{DD}. This ensures 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 cutoff. In practice, logic "1" 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 "0," the opposite biasing is used. A positive voltage is applied to the row (gate) and a negative voltage is applied to the column (drain), causing electrons to tunnel from the drain to the floating gate. This renders the threshold voltage of the FLOTOX device less than V_{DD}. When so 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, both the access transistor and the FLOTOX device are linear, bringing the column line to ground.

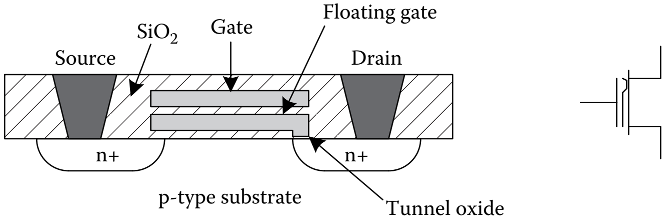


FIGURE 11.17
FLOTOX transistor.

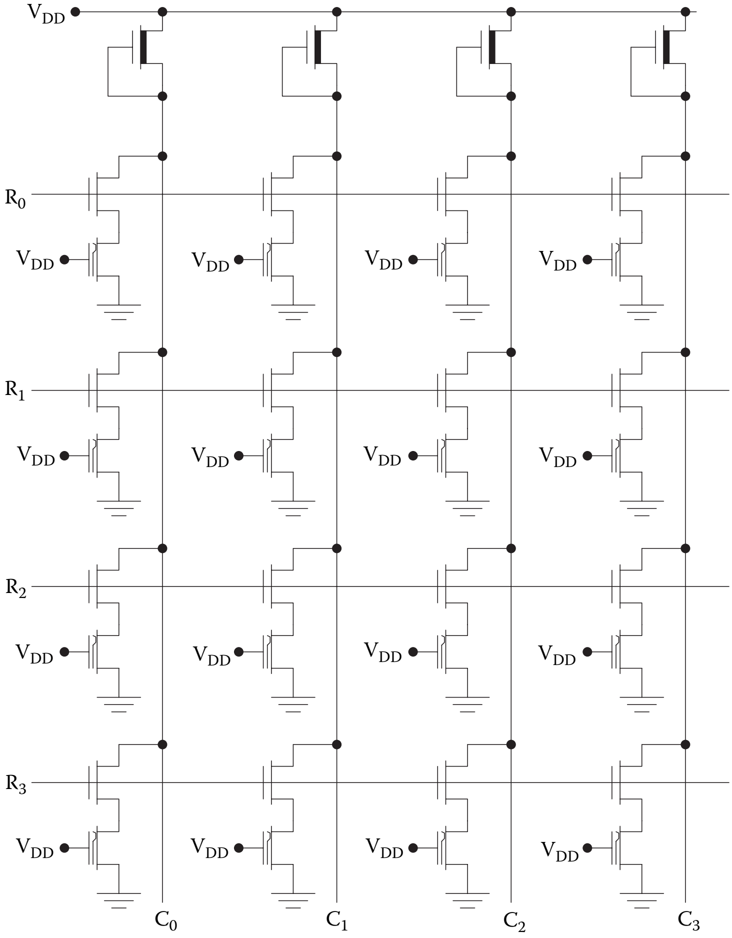


FIGURE 11.18
4 × 4 EEPROM.

The EEPROM needs two transistors per bit because it is not possible to tightly control the programmed threshold voltage of the FLOTOX transistor. After programming logic "0," 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.

11.8 Flash Memory

Flash memory [17–25] combines the flexibility of EEPROM with the high density of EPROM. This type of memory uses erase through oxide (ETOX) devices, which are similar to FAMOS transistors except that they have a thin tunneling oxide under the floating gate. A 4 × 4 NOR flash memory circuit is illustrated in Figure 11.19. The source line S is grounded for a read/write operation (see Figure 11.20), which effectively renders the circuitry in the same configuration as an NOR ROM.

For the NOR topology, programming is typically achieved by the avalanche injection of electrons, but NAND flash may use Fowler-Nordheim tunneling. Erasure is accomplished with Fowler-Nordheim tunneling in both NOR and NAND flash memories by applying a positive voltage on the source line S. This is similar to the case of the FLOTOX transistor, but an important difference is that erasure 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. This is why 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.

Significant innovations in flash memory include the replacement of the polysilicon floating gate with silicon nitride and the incorporation of dual gates. Both approaches have made it possible to store two bits per cell in

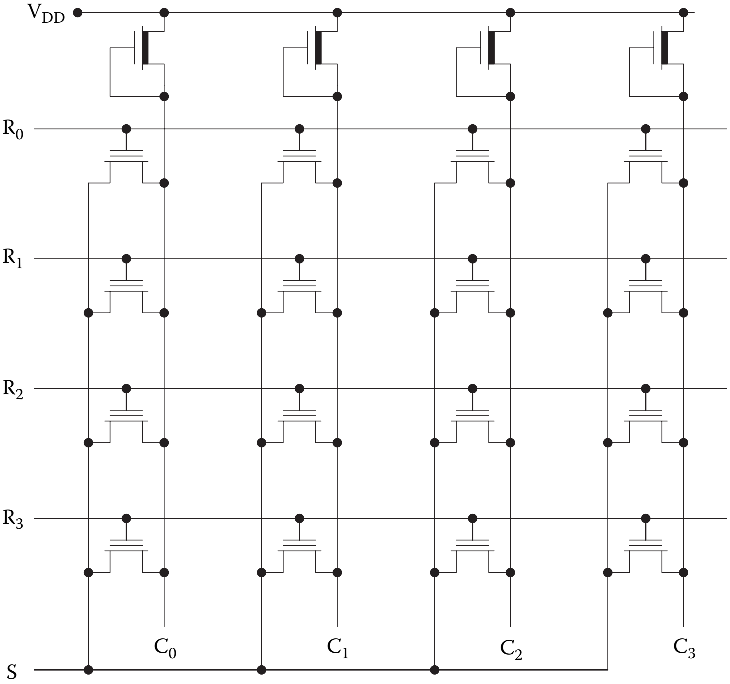


FIGURE 11.19
4 × 4 NOR flash memory.

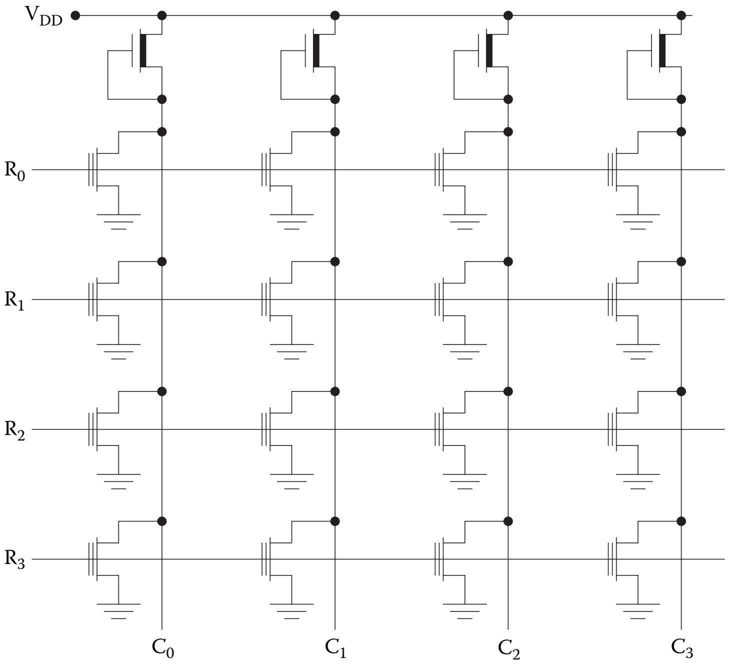


FIGURE 11.20
4 × 4 NOR flash memory with the S line grounded for a read operation.

commercial flash memory chips at the present time. Some high-density designs can store four (or more) bits per transistor, but the reduced voltage swings give rise to longer read times.

Figure 11.21 illustrates the NAND flash topology. The NAND flash requires fewer contacts, resulting in higher density than a NOR circuit and is therefore preferred for USB (universal serial bus) drives and digital camera cards. On the other hand, NAND flash tends to be slower because of the series-connected transistors.

11.9 Other Nonvolatile Memories

Flash memories are well suited to applications such as digital cameras, media players, and wireless phones, in which nonvolatile memory is necessary 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 ~10^6. This precludes the use of flash memory as core memory in

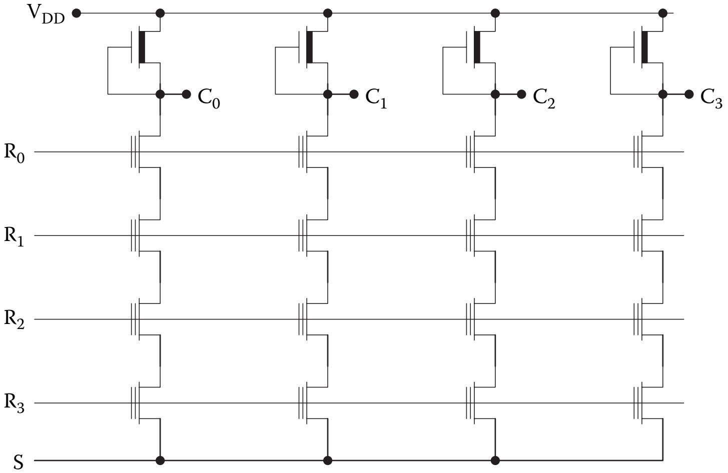


FIGURE 11.21
4 × 4 NAND flash memory.

digital computers, which would exceed this number of cycles in a short time. Instead, DRAM is used, but the drawback of DRAM is that the data are lost during power down. There is therefore a need for nonvolatile memory that can withstand unlimited erase/write cycles. Recently, new types of memories have emerged that satisfy these requirements. These are ferroelectric random access memory (FRAM) [26–34], magnetoresistive random access memory (MRAM) [35–37], and phase change memory (PCM) [38–41]. These technologies represent a major departure from today's digital memory devices because they use the properties of non-silicon materials.

In an FRAM, each bit is stored in a ferroelectric capacitor that is connected to an access transistor as shown in Figure 11.22. The capacitor is made using a ferroelectric material such as lead zirconium titanate. In the ferroelectric material, there is a built-in electric field that is determined by the polarity of electric domains in it. These domains can be flipped in polarity by the application of the appropriate bias voltage. Therefore, the two possible polarities of the domains can be used to represent "logic 1" and "logic 0."

To store a bit in the FRAM cell of Figure 11.23, 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 "0" or "1," respectively. The drive line R′ is driven in complementary manner to the bit line to provide the required bias polarity for the capacitor. This orients the domains in the ferroelectric capacitor such that a negative or positive voltage is stored in this capacitor to represent "0" or "1," respectively.

To read a bit from the FRAM cell, the word line R and the column line R are both 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

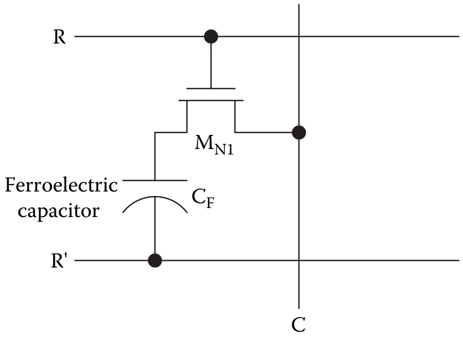


FIGURE 11.22
FRAM cell containing one transistor and one ferroelectric capacitor (1T1C cell).

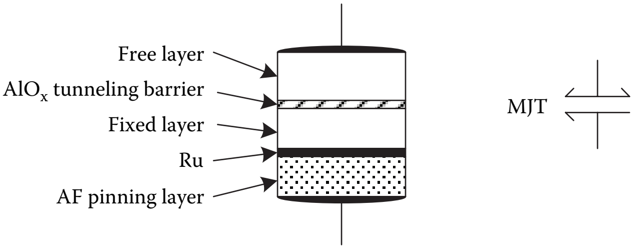


FIGURE 11.23
MTJ and circuit symbol.

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. The ferromagnetic material is made up 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 "1" or a "0." 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 (MJT). Figure 11.23 shows such an MJT 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 pinning layer such as iron-manganese or iridium-manganese with an intermediate layer of ruthenium. This combination creates a synthetic antiferromagnet. 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 11.24. 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 MJT parallel to that in the fixed layer. The application of the currents with opposite polarities fixes the magnetic field of the free layer in the antiparallel direction. Neither current alone is sufficient to program the free layer. Hence, 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.

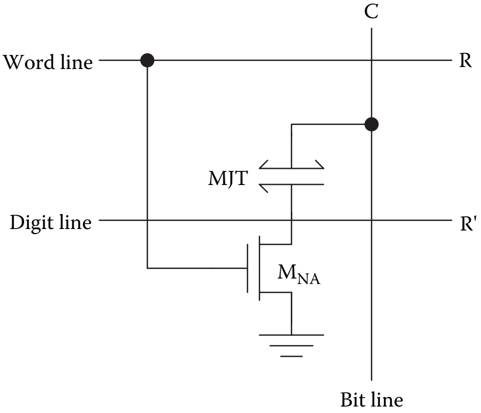


FIGURE 11.24
MRAM cell with one transistor and one magnetic tunneling junction (1T1MTJ cell).

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.

PCM, sometimes called ovonic unified memory, uses a chalcogenide alloy such as germanium-antimony-tellurium, the same type of material used in DVD R/W technology. The chalcogenide material can exist indefinitely in one of two phases: crystalline and amorphous. The amorphous phase is characterized by high resistivity, whereas the crystalline phase exhibits low resistivity. It is therefore possible to use one phase to represent logic "1" and the other phase to represent logic "0".* 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, allowing it to crystallize by a process of *solid-phase epitaxy*. To achieve this operation, each cell of the PCM comprises a programmable chalcogenide resistor, a resistance heater, and an isolation diode as shown in Figure 11.25.

Some of the properties of these nonvolatile memory technologies are summarized in Table 11.1. 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.

11.10 Access Times in Digital Memories

The speed of a memory circuit is usually specified in terms of the *access time*. The read access time is the delay from the time an address is presented at the address lines to the time data appear at the outputs of the integrated circuit. The write access time is the time required for a bit to be stored once the address is presented. In general, the read and write access times may be different.

* CD R/W and DVD R/W technology uses the dramatic difference in optical reflectivity for the two phases of the chalcogenide.

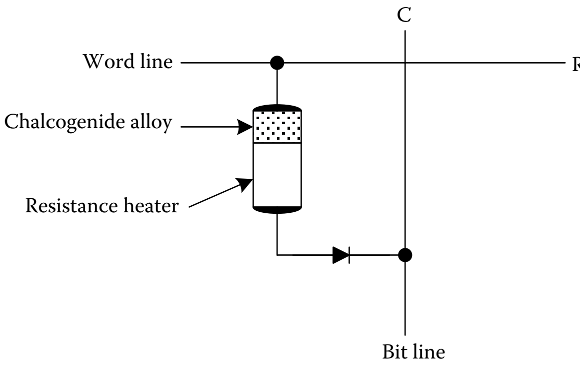


FIGURE 11.25
PCM cell.

Another important timing parameter is the cycle time. This is the reciprocal of the frequency at which addresses are presented to the memory circuit for read (or write) operations. Obviously, reliable operation requires that the cycle time be greater than the access time.

Usually, the access time in a digital memory is limited by the delay times for the interconnects forming the row and column lines, especially if these are formed by polysilicon or ion implanted regions of semiconductor. These materials have significantly higher specific resistivity than metal interconnect, and their associated 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 we must consider the delay associated with the entire length of the row line. Once

TABLE 11.1
Comparison of Nonvolatile Memories

Parameter	Flash	FRAM	MRAM	PCM
Maximum capacity (Mb)	256	64	1	4
Cell size*	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

* Normalized to the cell size for flash memory

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/fall time is negligible at the selected cell, we have to consider the delay associated with the column interconnect. 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, there is an additional delay 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. Therefore, 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:

$$\begin{aligned} t_{read} &= t_{word} + t_{bit} \\ &= \ln(2) R_w C_w \frac{2^N + 1}{2} + \ln(2) R_b C_b \frac{2^M + 1}{2} \\ &\approx \ln(2) R_w C_w 2^{N-1} + \ln(2) R_b C_b 2^{M-1}. \end{aligned} \tag{11.2}$$

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 manner 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, and this will increase the read time significantly. Such is the case for FRAM.

11.11 Row and Column Decoder Design

Any digital memory, arranged in 2^N rows × 2^M columns, must have decoders to select the correct word line and bit line based on N + M address bits.

The row decoder need only implement the necessary combinational logic to select the appropriate row line, by either driving it to V_{DD} (NOR array) or

0 (NAND array). Such a decoder circuit can be conveniently designed in the same style (NOR/NAND) as the memory array. For example, in a 4×4 NOR ROM, the row decoder must implement four functions of two address bits as shown in Table 11.2. (Active high word lines are assumed.) The desired logic functions can be implemented in a NOR array of transistors as shown in Figure 11.26.

To decode four word lines with a NAND array, assuming active low word lines, the four required logic functions are given in Table 11.3 and the NAND decoder realization is shown in Figure 11.27.

The column (bit line) decoder is not simply a combinational logic circuit but must transfer data (voltage levels) to and from the bit lines of the memory array. This function can be achieved using a binary tree of n-MOS pass transistors, and Figure 11.28 illustrates this for the case of four bit lines addressed by two column address bits B_0 and B_1.

11.12 Practical Perspective

For practical perspective articles, see the dynamic website at http://www.engr.uconn.edu/ece/books/ayers.

11.13 Summary

Digital memories store bits of information for later use by processors, displays, and input/output devices and are therefore crucial elements in most digital systems. All digital memory circuits use simple memory cells 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 store one or more bits.

TABLE 11.2

Row Decoder Logic for a Memory with Four Word Lines (rows) and Two Row Address Lines A_0 and A_1*

A_0	A_1	R_0	R_1	R_2	R_3
0	0	1	0	0	0
0	1	0	1	0	0
1	0	0	0	1	0
1	1	0	0	0	1

*The word lines are active high.

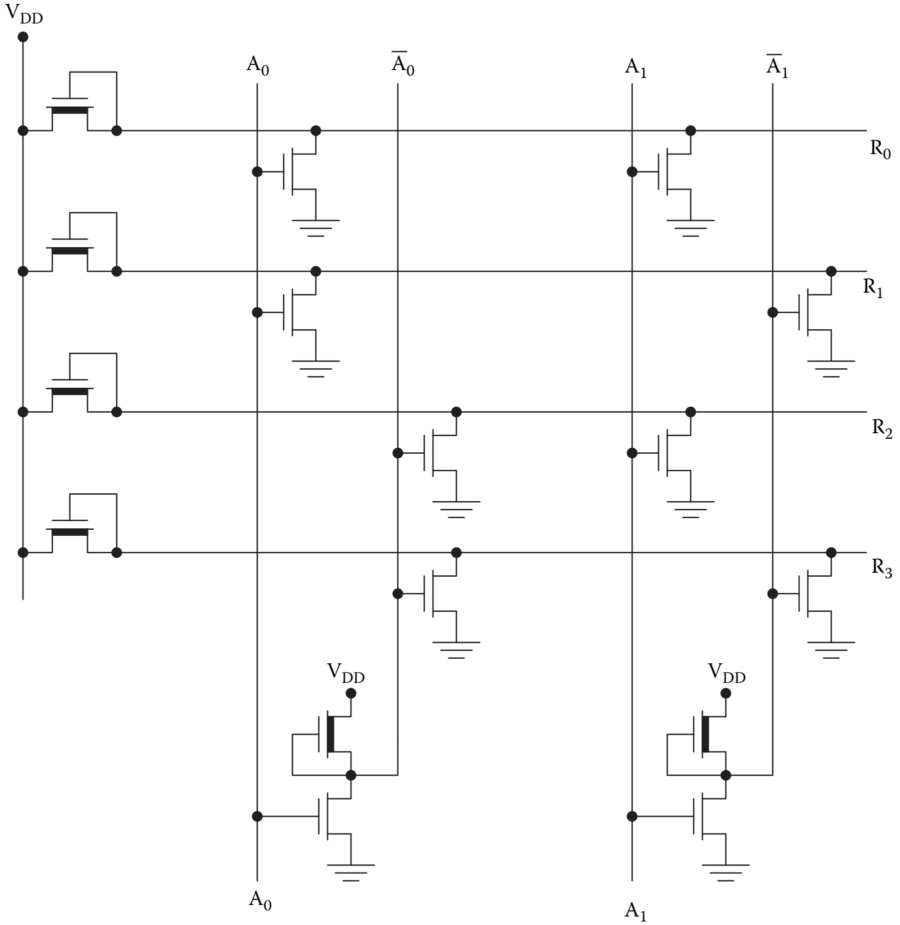


FIGURE 11.26
NMOS NOR row decoder with two row address lines and four word lines. (The word lines are active high. A_0 is the most significant bit.)

TABLE 11.3
Row Decoder Logic for a Memory with Four Word Lines (rows) and Two Row Address Lines A_0 and A_1*

A_0	A_1	$\overline{R_0}$	$\overline{R_1}$	$\overline{R_2}$	$\overline{R_3}$
0	0	0	1	1	1
0	1	1	0	1	1
1	0	1	1	0	1
1	1	1	1	1	0

*The word lines are active low.

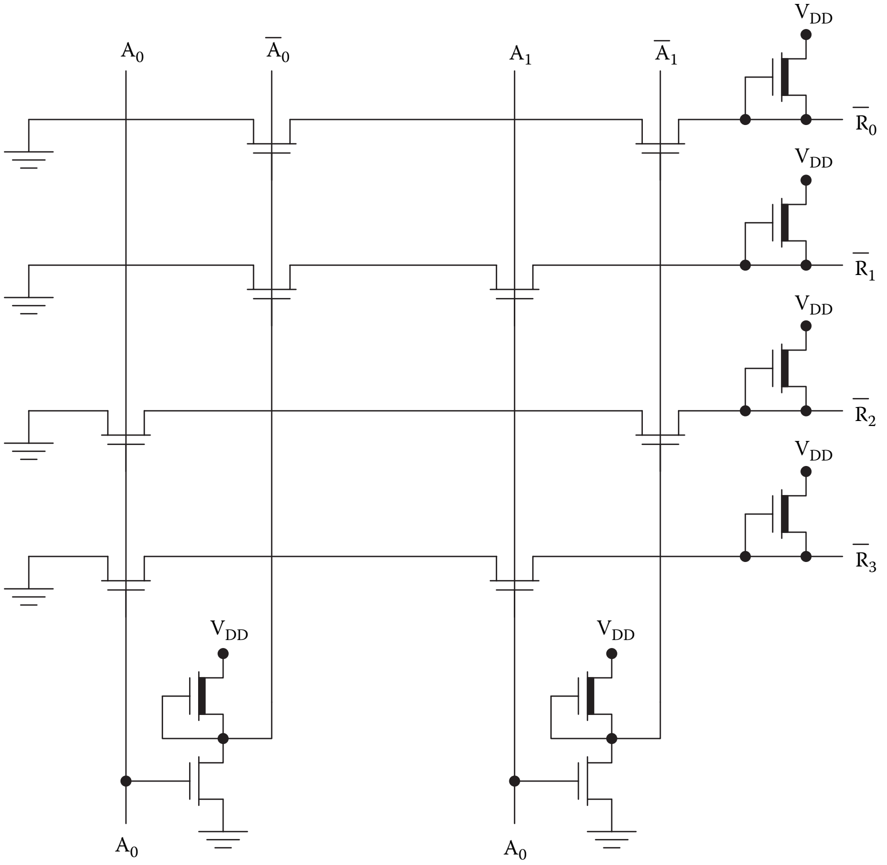


FIGURE 11.27
NMOS NAND row decoder with two row address lines and four word lines. (The word lines are active low. A_0 is the most significant bit.)

The data in any individual cell may be accessed by selecting the jth row and the kth 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 interpreted by the column decoder; in addition, the column decoding circuitry is responsible for transferring data in and out of the memory. For this reason, column lines are also known as *bit lines*. Row lines, however, select entire rows (or words) of data so these are called *word lines*.

Broadly speaking, memories may be classified as volatile or nonvolatile; volatile memories only retain their data while the system power is on. Volatile memories include SRAM and DRAM. Nonvolatile memories include ROM, PROM, EPROM, and EEPROM. Other nonvolatile memories that may

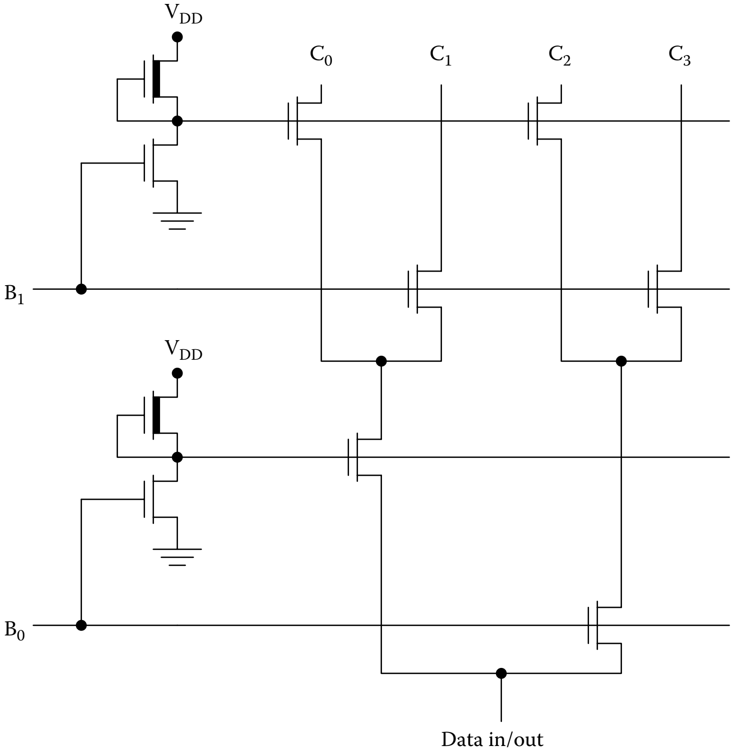


FIGURE 11.28
Binary tree column decoder for two column address bits B_0 and B_1 that address four bit lines C_0, C_1, C_2, and C_3. (B_0 is the most significant address bit.)

be erased and programmed electrically include flash memory, FRAM, MRAM, and PCM.

11.14 Exercises

E11.1. Suppose you are designing a memory chip which 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 100 Gb memory chip if each address holds 16 bits?

E11.2. What is the minimum number of pins required for a 1 Gb memory chip? (Include V_{DD}, GND, and CLK pins.)

E11.3. 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.

E11.4. Consider a 16 Mb DRAM with a square layout (4096 × 4096). 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 0.2 ns.

E11.5. 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.

For additional exercise problems, see the dynamic website at http://www.engr.uconn.edu/ece/books/ayers.

References

1. Fried, D.M., Hergenrother, J.M., Topol, A.W., Chang, L., Sekaric, L., Sleight, J.W., McNab, S.J., Newbury, J., Steen, S.E., Gibson, G., Zhang, Y., Fuller, N.C.M., Bucchignano, J., Lavoie, C., Cabral Jr., C., Canaperi, D., Dokumaci, O., Frank, D.J., Duch, E.A., Babich, I., Wong, K., Ott, J.A., Adams, C.D., Dalton, T.J., Nunes, R., Medeiros, D.R., Viswanathan, R., Ketchen, M., Jeong, M., Haensch, W., and Guarini, K.W., Aggressively scaled (0.143 μm^2) 6T-SRAM cell for the 32 nm node and beyond. *IEDM Tech. Dig.*, 261–264, 2004.
2. Witters, L., Collaert, N., Nackaerts, A., Demand, M., Demuynek, S., Delvaux, C., Lauwers, A., Baerts, M., Goodwin, M., Hendrickx, E., Heylen, N., Jaenen, P., Laidler, D., Leray, P., Locorotondo, S., Maenhoudt, M., Moclants, M., Pollentier, I., Ronse, K., Rooyackers, R., Van Aelst, J., Vandenberghe, G., Vandeweyer, T., Vanhaelemeersch, S., Van Hove, M., Van Olmen, J., Verhaegen, S., Verslujis, J., Vrancken, C., Wiaux, V., Willems, P., Wouters, J., Jurezak, M., and Biesemans, S., Integration of tall triple-gate devices with inserted-Ta_xN_y gate in a 0.274 μm^2 6T-SRAM cell and advanced CMOS logic circuits, Digest of Technical Papers. *2005 VLSI Technol.* 106–107, 2005.
3. Yang, F.-L., Huang, C.-C., Huang, C.-C., Chung, T.-X., Chen, H.-Y., Chang, C.-Y., Chen, H.-W., Lee, D.-H., Liu, S.-D., Chen, K.-H., Wen, C.-K., Cheng, S.-M., Yang, C.-T., Kung, L.-W., Lee, C.-L., Chou, Y.-J., Liang, F.-J., Shiu, L.-H., You, J.-W., Shu, K.-C., Chang, B.-C., Shin, J.-J., Chen, C.-K., Gau, T.-S., Wang, P.-W., Chan, B.-W., Hsu, P.-F., Shieh, J.-H., Fung, S.K.-H., Diaz, C.H., Wu, C.-M. M, See, Y.-C., Lin, B.J., Liang, M.-S., Sun, J.Y.-C., and Hu, C., 45 nm node planar-SOI technology with 0.296 μm^2 6T-SRAM cell, Digest of Technical Papers. *2004 Symp. VLSI Technol.*, 8–9, 2004.

4. Wicht, B., Nirschl, T., and Schmitt–Landsiedel, D., A yield-optimized latch-type sense amplifier. *Proc. 29th Eur. Solid-State Circuits Conf.*, 409-412, 2003.
5. Muller, K.P., Flietner, B., Hwang, C.L., Kleinhenz, R.L., Nakao, T., Ranade, R., Tsunashima, Y., Mii, T., Trench storage node technology for gigabit DRAM generations. *IEDM Tech. Dig.*, 507–510, 1996.
6. Kohyama, Y., Ozaki, T., Yoshida, S., Ishibashi, Y., Nitta, H., Inoue, S., Nakamura, K., Aoyama, T., Imai, K., and Hayasaka, N., A fully printable, self-aligned and planarized stacked capacitor DRAM cell technology for 1Gbit DRAM and beyond. *VLSI Technol. Tech. Dig.*, 17–18, 1997.
7. Kim, K., Hwang, C.-G., and Lee, J., DRAM technology perspective for gigabit era. *IEEE Trans. Electron Dev.*, 45, 598, 1998.
8. Parke, S.A., Optimization of DRAM sense amplifiers for the gigabit era. *Proc. 40th Midwest Symp. Circuits Syst.*, 209–212, 1997.
9. Lu, N.C.C., and Chao, H.H., Half-V_{DD} bit-line sensing scheme in CMOS DRAMs. *IEEE J. Solid-State Circuits*, 19, 451–454, 1984.
10. Natori, K., Sensitivity of dynamic MOS flip-flop sense amplifiers. *IEEE Tran. Electron Dev.*, 33, 482–454, 1986.
11. Dhong, S., Lu, N.C.C., Hwang, W., and Parke, S., High-speed sensing scheme for CMOS DRAMs. *IEEE J. Solid-State Circuits*, 23, 34–40, 1988.
12. Okhonin, S., Nagoga, M., Sallese, J.M., and Fazan, P., A capacitor-less 1T-DRAM cell. *IEEE Electron Dev. Lett.*, 23, 85–87, 2002.
13. Jeong, H., Song, K.-W., Park, I.H., Kim, T.-H., Lee, Y.S., Kim, S.-G., Seo, J., Cho, K., Lee, K., Shin, H., Lee, J.D., and Park, B.-G., A new capacitorless 1T DRAM cell: surrounding gate MOSFET with vertical channel (SGVC cell). *IEEE Trans. Nanotechnol.*, 6, 352–357, 2007.
14. Cui, W., Design of small area and low power consumption mask ROM. *IEEE Int. Conf. Integrated Circuit Design Technol.*, 1–4, 2007.
15. Frahman-Bentchkowsky, D., A fully decoded 2048-bit electrically programmable FAMOS read-only memory. *IEEE J. Solid-State Circuits*, 6, 301–306, 1971.
16. Meng, M., and Noren, K.V., A macromodel for FLOTOX EEPROM. *IEEE 39th Midwest Symp. Circuits Syst.*, 39–42, 1997.
17. Pavan, P., Bez, R., Olivo, P., and Zanoni, E., Flash memory cells: an overview. *Proc. IEEE*, 85, 1248–1271, 1997.
18. Barre, A.G., Flash memory: an exploding alternative to fixed hard disks. *Dig. 1993 Int. Magnetics Conf.*, BZ-06, 1993.
19. Masuoka, F., and Endoh, T., Flash memories, their status and trends. *Proc. 4th Int. Conf. Solid-State IC Technol.*, 128, 1995.
20. Aritome, S., Advanced flash memory technology and trends for file storage application, Technical Digest. *2000 Int. Electron Dev. Meet.*, 763, 2000.
21. Wett, T., and Levy, S., Flash–the memory technology of the future that's here today. *Proc. 1995 IEEE Nat. Aerospace Electron. Conf.*, 359, 1995.
22. Lai, S., Flash memories: where we were and where we are going, Technical Digest. *1998 Int. Electron Dev. Meet.*, 971, 1998.
23. Lorenzini, M., Rudan, M.V., Baccarani, G., A dual gate flash EEPROM cell with two-bit storage capacity components. *IEEE Trans. Packaging Manufact. Technol. Part A*, 20, 182, 1997.

24. 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.
25. 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, Digest of Technical Papers. *1999 IEEE Int. Solid-State Circuits Conf.*, 112, 1999.
26. Ramtron Corporation, http://www.ramtron.com.
27. Fujitsu Corporation, http://www.fujitsu.com.
28. Chung, Y., Experimental 128-kbit ferroelectric memory with 10^{12} endurance and 10-year data retention. *IEEE Proc. Circuits Dev. Syst.*, 149, 136, 2002.
29. 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, Digest of Technical Papers. *2002 Symp. VLSI Technol.*, 210, 2002.
30. 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, Digest of Technical Papers. *2002 IEEE Int. Solid-State Circuits Conf.*, 162, 2002.
31. 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, Digest of Technical Papers. *2000 Symp. VLSI Technol.*, 34, 2000.
32. 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, Digest of Technical Papers. *1999 Symp. VLSI Technol.*, 141, 1999.
33. 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, Digest of Technical Papers. *1999 IEEE Int. Solid-State Circuits Conf.*, 104, 1999.
34. 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, Digest of Technical Papers. *1998 Symp. VLSI Technol.*, 126, 1998.
35. 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, Digest of Technical Papers. *2002 Symp. VLSI Circuits*, 158, 2002.
36. Naji, P.K., Durlam, M., Tehrani, S., Calder, J., and DeHerrera, M.F., A 256 kb 3.0 V 1T1MTJ nonvolatile magnetoresistive RAM, Digest of Technical Papers. *2001 IEEE Int. Solid-State Circuits Conf.*, 122, 2001.
37. 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 Mem. Technol. Conf.*, 43, 1998.

38. Ovonyx Inc., http://www.ovonyx.com.
39. Gill, M., Lowrey, T., and Park, J., Ovonic unified memory: a high-performance nonvolatile memory technology for stand-alone memory and embedded applications, Digest of Technical Papers. *2002 IEEE Int. Solid-State Circuits Conf.*, 202, 2002.
40. Lai, S., and Lowrey, T., OUM: a 180 nm nonvolatile memory cell element technology for stand alone and embedded applications, Technical Digest. *2001 Int. Electron Dev. Meet.*, 36.5.1, 2001.
41. Maimon, J., Spall, E., Quinn, R., and Schnur, S., Chalcogenide-based nonvolatile memory technology. *Proc. 2001 IEEE Aerospace Conf.*, 2289, 2001.

12

Input/Output and Interface Circuits

12.1 Introduction

Thus far, the emphasis of this book has been the design and analysis of gates, latches, flip-flops, and memory elements. However, digital integrated circuits require the implementation of special input, output, and interface circuits as well. Inputs must have protection circuitry to prevent ESD damage during handling as well as transmission gates for enable/disable operation. Output pins generally require high-current output drivers and tri-state operation to allow compatibility with busses. Interface circuits are needed for voltage level shifting between circuits operating with different voltages.

12.2 Input Electrostatic Discharge Protection

Input pins to an integrated circuit may be subjected to electrostatic discharge from people, other components, or equipment carrying thousands of volts. These ESD events may occur during integrated circuit manufacture, handling, or product assembly. Personnel develop electrostatic charges of up to ~1 μC by casual contact with carpets, seats, and desks; subsequent handling of an integrated circuit can give rise to an ESD event that may resemble the discharge of a ~100 pF capacitor, charged to ~1.5 kV, through a ~1.5 kΩ resistance. Such a discharge to an unprotected CMOS circuit can give rise to a number of device failure mechanisms, including destructive oxide breakdown, polysilicon melting and filamentation, and metal contact spiking or filamentation [1]. It is therefore necessary to include an ESD protection network in line with each integrated circuit input pin.

A simple ESD protection network is illustrated in Figure 12.1. The silicon p-n diodes D_1 and D_2 clamp the input voltage such that

$$-V_D \leq V'_{IN} \leq \left(V_{DD} + V_D\right), \tag{12.1}$$

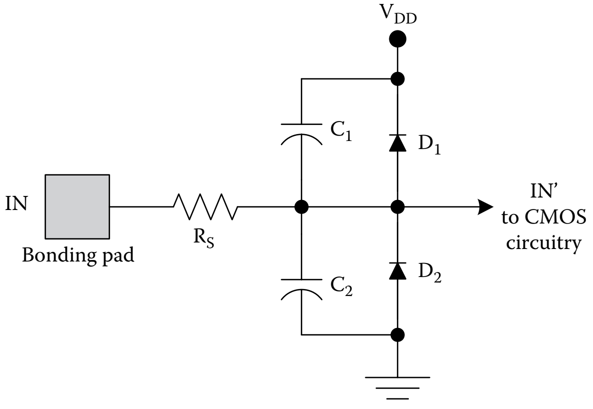


FIGURE 12.1
Simple ESD protection network.

where V_D is the diode turn-on voltage (approximately 0.7 V for silicon p-n junctions). The series resistance R_s appearing in the diffused silicon regions assumes the difference $V_{IN} - V'_{IN}$ and is therefore a necessary part of the protection network. The capacitances C_1 and C_2 are associated with the p-n junction diodes and, together with R_s, form an integrator (low-pass filter) that reduces the amplitude of the input voltage apart from the clamping action of the diodes. However, this low-pass filter also limits the rise time for the input signal.

There are many variations on the ESD protection network shown in Figure 12.1 [1–3]. For example, the diodes may be replaced with thyristors or diode-connected bipolar transistors. All of these bipolar devices are relatively large, and this has led to the development of ESD protection networks that can be fabricated *underneath* the bonding pads to conserve die area.

12.3 Input Enable Circuits

It is often desirable to be able to isolate an input from a bus when data are not being read, and this can be achieved by the insertion of a transmission gate as shown in Figure 12.2.

A CMOS transmission gate may be constructed using complementary n-MOS and p-MOS pass transistors as illustrated in Figure 12.3 [4]. Because this simple circuit requires complementary *ENABLE* and $\overline{ENABLE}$ signals, it is common to include an inverter as shown in Figure 12.4.

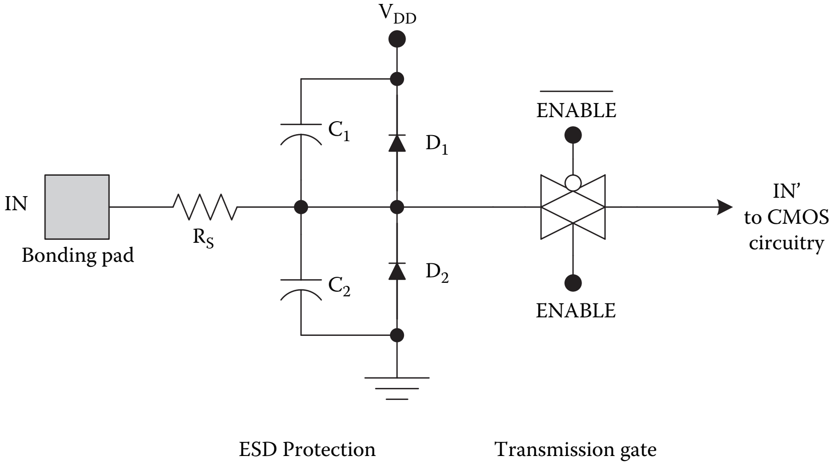


FIGURE 12.2
Input circuitry including an ESD protection network and a transmission gate.

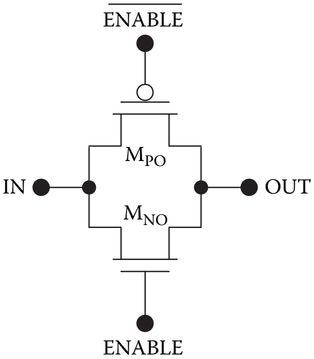


FIGURE 12.3
CMOS transmission gate with complementary enable inputs.

The CMOS transmission gates shown in Figures 12.3 and 12.4 are bidirectional. That is, because the MOS transistors are symmetric, current can flow from IN to OUT or vice versa in the enabled gate with no change in the "on" resistance. However, the "on" resistance is a function of the input voltage.

12.3.1 CMOS Transmission Gate

Consider the operation of the four-transistor circuit shown in Figure 12.4. If $V_{ENABLE} = 0$, then the voltage at the gate of the p-MOS transistor will be V_{DD}. For the n-MOS transistor, $V_{GSN} = -V_{IN}$ so this device will be cutoff for

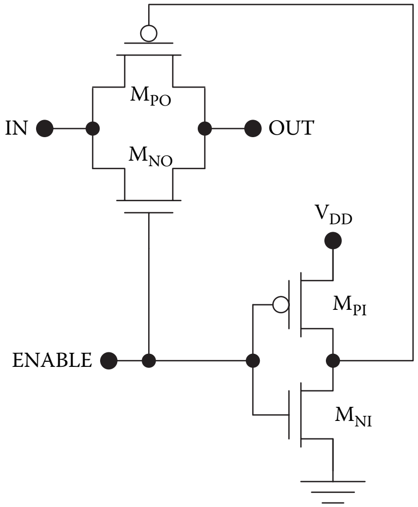


FIGURE 12.4
CMOS transmission gate.

any input voltage in the range $0 \le V_{IN} \le V_{DD}$. The p-MOS transistor will also be cutoff for any input in this range because $V_{GSN} = V_{DD} - V_{IN}$. Therefore, with $V_{ENABLE} = 0$ the transmission gate is in the "off" state, isolating the input and output nodes from one another.

On the other hand, if $V_{ENABLE} = V_{DD}$, the transmission gate is enabled and passes the voltage at the input to the output node. To analyze this situation, we will assume quasi-static conditions with a sufficiently small current so that $V_{OUT} \approx V_{IN}$, that is, $V_{DSN} \approx -V_{DSP} \approx 0$. For the n-MOS transistor, $V_{GSN} = V_{DD} - V_{IN}$ so this device will be linear if $V_{IN} \le V_{DD} - V_{TN}$ but cutoff if $V_{IN} \ge V_{DD} - V_{TN}$. For the p-MOS transistor, $V_{GSP} = -V_{IN}$ so this device will be linear if $V_{IN} \ge -V_{TP}$ but cutoff if $V_{IN} \le -V_{TP}$. This results in three regimes for the operation of the transmission gate as shown in Table 12.1.

TABLE 12.1
Regimes of Operation for the Enabled CMOS Transmission Gate

Regime	Voltage conditions	n-MOS mode	p-MOS mode
1	$V_{IN} \le -V_{TP}$	Linear	Cutoff
2	$-V_{TP} \le V_{IN} \le (V_{DD} - V_{TN})$	Linear	Linear
3	$(V_{DD} - V_{TN}) \le V_{IN}$	Cutoff	Linear

12.3.1.1 Regime One: n-MOS Linear and p-MOS Cutoff

In operation regime one, the n-MOS is linear, whereas the p-MOS is cutoff. For the linear n-MOS transistor,

$$I_D = K_N\left[\left(V_{DD} - V_{IN} - V_{TN}\right)V_{DS} - V_{DS}^2/2\right]. \tag{12.2}$$

The "on" resistance for the n-MOS transistor is

$$R_{ONN} = \left(\left.\frac{\partial I_D}{\partial V_{DS}}\right|_{V_{DS}=0}\right)^{-1} = \frac{1}{K_N\left(V_{DD} - V_{IN} - V_{TN}\right)}. \tag{12.3}$$

Although the p-MOS device is cutoff, the overall "on" resistance for the transmission gate is equal to that for the n-MOS transistor alone:

$$R_{ON} = \frac{1}{K_N\left(V_{DD} - V_{IN} - V_{TN}\right)}. \tag{12.4}$$

12.3.1.2 Regime Two: n-MOS Linear and p-MOS Linear

In operation regime two, both transistors are linear. Following the same line of reasoning as above, the individual "on" resistances are

$$R_{ONN} = \frac{1}{K_N\left(V_{DD} - V_{IN} - V_{TN}\right)} \tag{12.5}$$

and

$$R_{ONP} = \frac{1}{K_P\left(V_{IN} + V_{TP}\right)}. \tag{12.6}$$

These drain resistances act in parallel so the overall "on" resistance for the transmission gate is

$$R_{ON} = \left(\frac{1}{R_{ONN}} + \frac{1}{R_{ONP}}\right)^{-1} = \left[K_N\left(V_{DD} - V_{IN} - V_{TN}\right) + K_P\left(V_{IN} + V_{TP}\right)\right]^{-1}. \tag{12.7}$$

12.3.1.3 Regime Three: n-MOS Cutoff and p-MOS Linear

In operation regime three, only the p-MOS transistor is conducting so

$$R_{ON} = \frac{1}{K_P\left(V_{IN} + V_{TP}\right)}. \tag{12.8}$$

12.3.1.4 Overall Characteristic of CMOS Transmission Gate

The overall "on" resistance characteristic, with $V_{ENABLE} = V_{DD}$, is

$$R_{ON} = \begin{cases} \left[K_N\left(V_{DD} - V_{IN} - V_{TN}\right)\right]^{-1}; & 0 \le V_{IN} \le V_{TN} \\ \left[K_N\left(V_{DD} - V_{IN} - V_{TN}\right) + K_P\left(V_{IN} + V_{TP}\right)\right]^{-1}; & V_{TN} \le V_{IN} \le \left(V_{DD} - V_{TP}\right) \\ \left[K_P\left(V_{IN} + V_{TP}\right)\right]^{-1} & \left(V_{DD} + V_{TP}\right) \le V_{IN} \le V_{DD} \end{cases}. \quad (12.9)$$

In the special case of a symmetric transmission gate, for which $K_N = K_P$ and $V_{TN} = |V_{TP}|$, the "on" resistance is constant throughout regime two.

Example 12.1 Asymmetric CMOS Transmission Gate

Calculate the "on" resistance for the CMOS transmission gate of Figure 12.5 as a function of V_{IN} and with $V_{ENABLE} = 2.5V$.

Solution: The "on" resistance is given by

$$R_{ON} = \begin{cases} \left[400\mu A/V^2\left(2.0V - V_{IN}\right)\right]^{-1}; & (\text{regime one})\, 0 \le V_{IN} \le 0.5V \\ \left[400\mu A/V^2\left(2.0V - V_{IN}\right) + 300\mu A/V^2\left(V_{IN} - 0.5V\right)\right]^{-1}; & (\text{regime two})\, 0.5V \le V_{IN} \le 2.0V \\ \left[300\mu A/V^2\left(V_{IN} - 0.5V\right)\right]^{-1} & (\text{regime three})\, 2.0V \le V_{IN} \le 2.5V \end{cases}$$

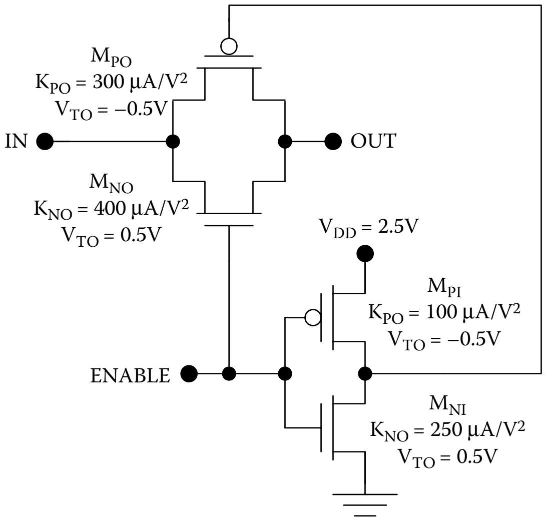


FIGURE 12.5
Example CMOS transmission gate.

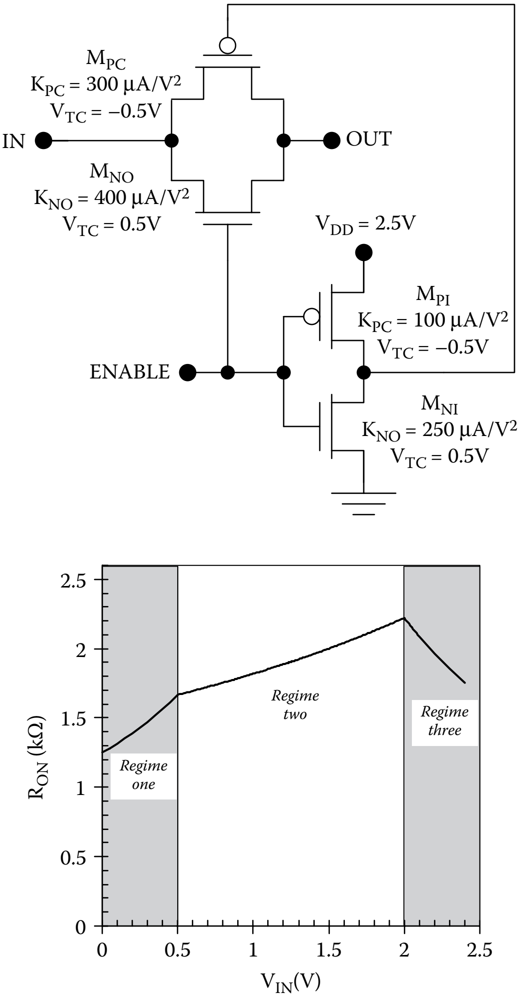


FIGURE 12.6
Calculated "on" resistance characteristic for an asymmetric CMOS transmission gate with $K_{PO} = 300\ \mu A/V^2$, $V_{TP} = -0.5$ V, $K_{NO} = 400\ \mu A/V^2$, $V_{TN} = 0.5$ V, and $V_{DD} = 2.5$ V.

This characteristic is plotted in Figure 12.6. The maximum "on" resistance is obtained with $V_{IN} = (V_{DD} + V_{TP})$ because the p-MOS transistor is weaker in this asymmetric circuit.

Example 12.2 Symmetric CMOS Transmission Gate

Calculate the "on" resistance characteristic for the transmission gate of Figure 12.7 with $V_{ENABLE} = 2.5V$. (In a symmetric CMOS transmission gate such as this,

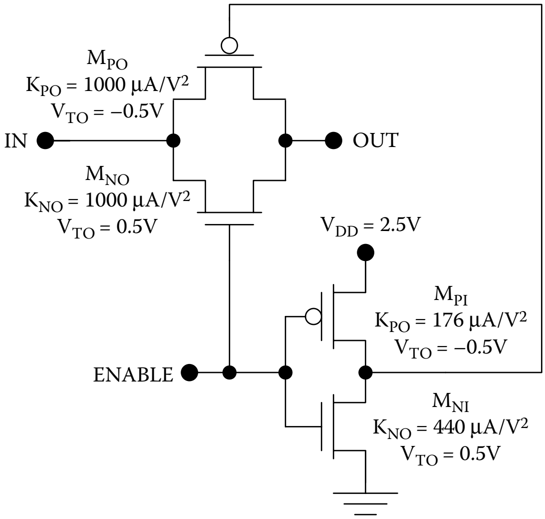


FIGURE 12.7
Example symmetric CMOS transmission gate.

$K_{NO} = K_{PO}$ and $V_{TN} = |V_{TP}|$, but it is not necessary for the inverter to have matched transistors.)

Solution: The "on" resistance is given by

$$R_{ON} = \begin{cases} \left[1mA/V^2\left(2.0V - V_{IN}\right)\right]^{-1}; & \text{(regime one)}\, 0 \le V_{IN} \le 0.5V \\ 667\Omega; & \text{(regime two)}\, 0.5V \le V_{IN} \le 2.0V \\ \left[1mA/V^2\left(V_{IN} - 0.5V\right)\right]^{-1} & \text{(regime three)}\, 2.0V \le V_{IN} \le 2.5V \end{cases} .$$

The "on" resistance is constant in regime two, as shown in Figure 12.8.

12.4 CMOS Output Buffers

A modern VLSI CMOS chip contains 1 million or more gates but only ~10^3 external connections. Therefore, most of the gates experience only on-chip loads (~10 fF) and are capable of propagation delays in picoseconds without the need for buffering. On the other hand, gates driving output pins

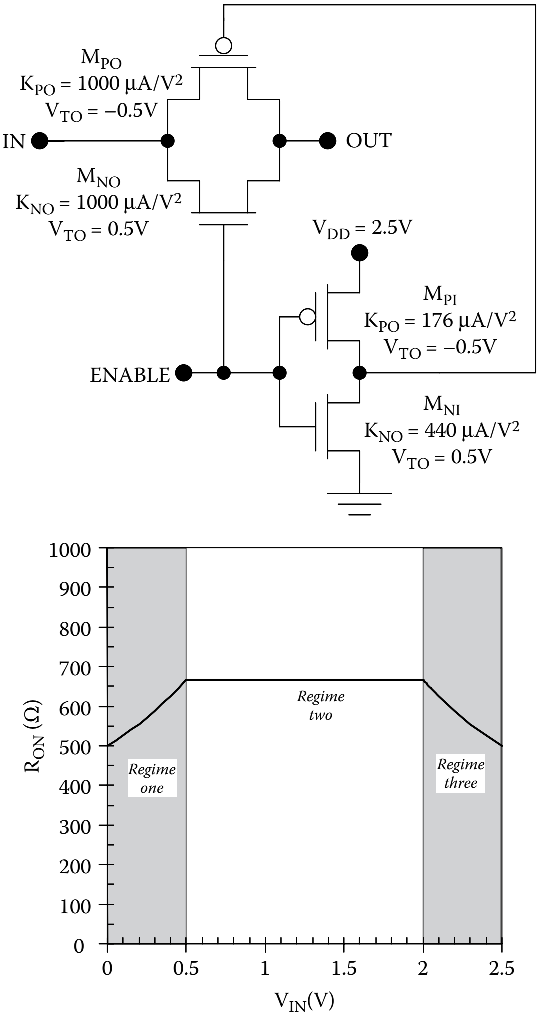


FIGURE 12.8
Calculated "on" resistance characteristic for a symmetric CMOS transmission gate with $K_{PO} = K_{NO} = 1\ \text{mA/V}^2$.

experience significantly greater load capacitances (~10 pF) so buffering is necessary to achieve acceptable off-chip data rates [5].

Figure 12.9 illustrates a CMOS inverter with N stages of inverting buffers placed between it and a load capacitance C_L. The device widths are scaled progressively by a factor of k at each stage. To analyze the overall behavior of this buffer arrangement, we will assume that all of the inverter stages are

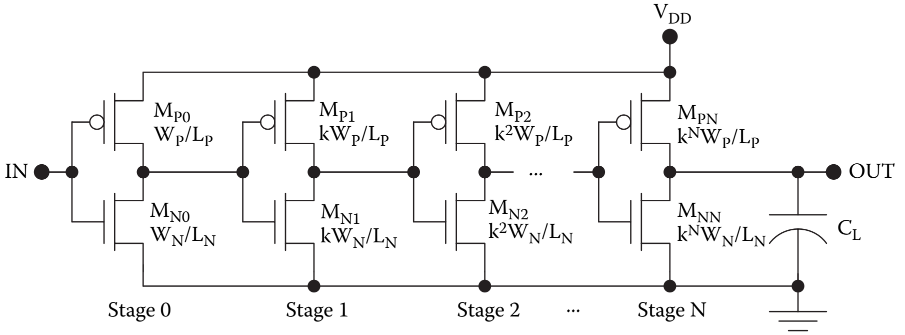


FIGURE 12.9
A CMOS inverter driving a load capacitance using N scaled buffer stages.

symmetric, i.e. for all stages $V_{TN} = |V_{TP}| = V_T$ and $\Gamma_N = \Gamma_P = \Gamma$, and for the individual stages, $K_{N0} = K_{P0} = K_0$, $K_{N1} = K_{P1} = K_1$, …, $K_{NN} = K_{PN} = K_N$.

The overall propagation delay is the sum of the individual propagations delays for the cascaded stages:

$$t_P = \sum_{j=0}^{N} t_{Pj}\,. \tag{12.10}$$

For stages 0 through N − 1, we will assume that the load capacitance is equal to the input capacitance of the next stage. Thus,

$$\begin{aligned} t_{Pj} &= \frac{C_{Lj}}{K_j}\Gamma = \frac{k^j C_{IN0}}{k^{j-1}K_0}\Gamma = \frac{kC_{ox}\left(1+\mu_n/\mu_p\right)\left[W_{N0}L_{N0}+2W_{N0}L_{OV}\right]}{\left(W_{N0}/L_{N0}\right)\left(\mu_n C_{ox}\right)}\Gamma \\ &= k\left(1+\mu_n/\mu_p\right)\left[L_{N0}^2+2L_{OV}L_{N0}\right]\Gamma/\mu_n;\quad 0 \le j \le (N-1), \end{aligned} \tag{12.11}$$

and each of these stages exhibits the same propagation delay. For the final stage,

$$t_{PN} = \frac{C_L}{K_N}\Gamma = \frac{C_L}{k^N K_0}, \tag{12.12}$$

The overall propagation delay is therefore

$$t_P = Nk\left(1+\mu_n/\mu_p\right)\left[L_{N0}^2+2L_{OV}L_{N0}\right]\Gamma/\mu_n + \frac{C_L}{k^N K_0}\Gamma\,. \tag{12.13}$$

The first term, as a result of stages 0 through N − 1, increases with k but the second term, as a result of the final stage, decreases with increasing k. There is thus an optimum scale factor k $(k_{OPT} = e \approx 2.7)$ that minimizes the overall propagation delay for a given load capacitance. However, this value of k is *not* optimum in terms of the silicon area consumed, and so larger scale factors may be used in practical output buffers.

Example 12.3 Unbuffered CMOS Propagation Delay

Estimate the propagation delay for a single CMOS stage driving a 10 pF off-chip load as shown in Figure 12.10.

Solution: The process transconductance parameters for this technology are

$$k_P^{'} = \frac{\mu_p \varepsilon_{ox}}{t_{ox}} = \frac{(230cm^2 / Vs)(3.9)(8.85 \times 10^{-14} F / cm)}{9 \times 10^{-7} cm} = 88 \mu A / V^2$$

and

$$k_N^{'} = \frac{\mu_n \varepsilon_{ox}}{t_{ox}} = \frac{(580cm^2 / Vs)(3.9)(8.85 \times 10^{-14} F / cm)}{9 \times 10^{-7} cm}$$

$$= 220 \mu A / V^2.$$

The device K values are

$$K_P = k_P^{'} \frac{W_P}{L_P} = 88 \mu A / V^2 \left(\frac{3.0}{0.6} \right) = 440 \mu A / V^2$$

and

$$K_N = k_N^{'} \frac{W_N}{L_N} = 220 \mu A / V^2 \left(\frac{1.2}{0.6} \right) = 440 \mu A / V^2 .$$

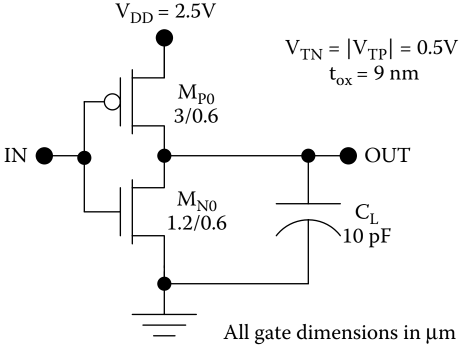


FIGURE 12.10
Unbuffered CMOS inverter stage driving an off-chip load of 10 pF.

The propagation delay parameter is

$$\begin{aligned}\Gamma &= \frac{1}{(V_{DD}-V_T)}\left[\frac{2V_T}{(V_{DD}-V_T)}+\ln\left(\frac{3V_{DD}-4V_T}{V_{DD}}\right)\right]\\ &= \frac{1}{(2.5V-0.5V)}\left[\frac{2(0.6V)}{(2.5V-0.5V)}+\ln\left(\frac{3(2.5V)-4(0.5V)}{2.5V}\right)\right]\\ &= 0.69V^{-1}.\end{aligned}$$

Therefore, the propagation delay is

$$t_P \approx \frac{10\times10^{-12}F}{440\mu A/V^2}0.69V^{-1} = 15.7ns\ .$$

Example 12.4 Propagation Delay with Two Buffer Stages

Estimate the propagation delay the CMOS inverter with two buffer stages driving a 10 pF load as depicted in Figure 12.11. (The scale factor for the buffers is five as shown.)

Solution: As in the previous example, $\Gamma = 0.69V^{-1}$ and $K_{P0} = K_{N0} = 440\mu A/V^2$. The oxide capacitance per unit area is $C_{ox} = \varepsilon_{ox}/t_{ox} = 3.83\times10^{-15}F/\mu m^2$, and the load capacitance experienced by the first stage is

$$\begin{aligned}C_{L0} = C_{IN1} &= 3.83\times10^{-15}F/\mu m^2\left[(15)(0.6)+2(15)(0.1)\right]\\ &\quad + 3.83\times10^{-15}F/\mu m^2\left[(6)(0.6)+2(6)(0.1)\right]\\ &= 64fF\end{aligned}$$

The propagation delay for the zeroeth stage is

$$t_{P0} \approx \frac{64\times10^{-15}F}{440\mu A/V^2}0.69V^{-1} = 100ps\ .$$

For the first buffer stage, both transistors have been scaled up in width by a factor of five, so the device transconductance parameter is five times larger:

$$K_1 = 5K_0 = 2200\mu A/V^2\ .$$

The load capacitance is equal to the input capacitance for stage two, and this is also increased fivefold:

$$C_{L1} = C_{IN2} = 5C_{IN1} = 320fF\ .$$

Therefore, the propagation delay for the first buffer stage is the same as for the zeroeth stage:

$$t_{P1} \approx \frac{320\times10^{-15}F}{2200\mu A/V^2}0.69V^{-1} = 100ps.$$

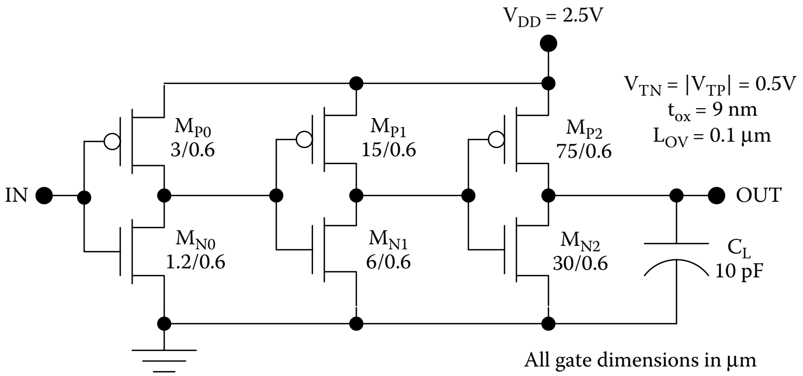


FIGURE 12.11
Unbuffered CMOS inverter stage driving an off-chip load of 10 pF.

For the second buffer stage, the device transconductance parameters are

$$K_2 = 5K_1 = 11mA/V^2,$$

whereas the load capacitance is 10 pF (the external load). The propagation delay for this stage is

$$t_{P2} \approx \frac{10\times10^{-12}F}{11mA/V^2}0.69V^{-1} = 630ps.$$

The overall delay is

$$t_P = t_{P0} + t_{P1} + t_{P2} = 100ps + 100ps + 630ps = 830ps.$$

Example 12.5 Propagation Delay with Four Buffer Stages

Estimate the propagation delay the CMOS inverter with four buffer stages driving a 10 pF load as illustrated in Figure 12.12.

Solution: Here, $N = 4$ and $k = 5$. $\Gamma = 0.69V^{-1}$ and $K_0 = 440\mu A/V^2$. The overall propagation delay for the five cascaded inverters is

$$t_P = Nk\left(1+\mu_n/\mu_p\right)\left[L_{N0}^2 + 2L_{OV}L_{N0}\right]\Gamma/\mu_n + \frac{C_L}{k^N K_0}\Gamma$$

$$= 4(5)(1+580/230)\left[\left(0.6\times10^{-4}cm\right)^2 + 2\left(0.1\times10^{-4}cm\right)\left(0.6\times10^{-4}cm\right)\right]$$

$$\left(0.69V^{-1}\right)/580cm^2/Vs + \frac{10\times10^{-12}F\left(0.69V^{-1}\right)}{5^4\left(440\mu A/V^2\right)}$$

$$= 4(100ps) + 25ps = 425ps.$$

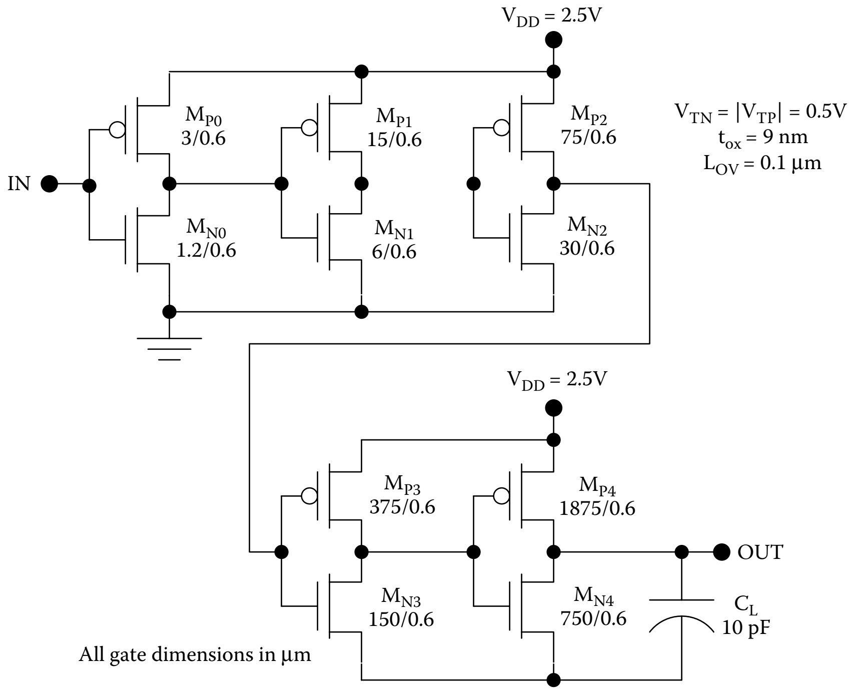


FIGURE 12.12
Unbuffered CMOS inverter stage driving an off-chip load of 10 pF.

Although the overall propagation delay has been reduced significantly compared with the unbuffered example, considerable chip area is taken up by the buffer stages. The last stage alone consumes ~5^4 times the silicon area of the zeroeth stage. In fact, this design is not optimized with respect to either the silicon area or the overall delay. Generally, in an optimum design, all stages have roughly equal delays. Therefore, too many stages have been used in the present example.

12.5 Tri-State Outputs

Many digital systems route signals on data busses, address busses, and control busses. In such a system, multiple driving devices share the common copper lines of a bus and to avoid potential conflicts only one of the devices should attempt to drive the bus at a particular time. This is achieved by tri-state outputs, in which the third state is the high impedance or "high Z" state. In the high Z state, the output is effectively disabled so that this electrical node may float to whatever voltage is set by another (enabled) device. Tri-state operation can be implemented by placing a transmission gate after a two-state device or by making internal modifications to the two-state gate; however, the latter approach is far more common.

Tri-state function can be provided in any CMOS gate by the addition of four MOSFETs, as shown in Figure 12.13 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, both M_{NE} and M_{PE} are cut-off, providing high Z operation. If the enable is high, both 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 12.14. Here too, only four extra MOSFETs are required to provide the tri-state function. This concept can be extended to CMOS gates with any level of complexity. However, due consideration must be given to scaling of the MOSFETs. The same design rules discussed in Chapter 6 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 12.15. This design requires a total of 12 MOS transistors but only two output driver transistors. Although the tri-state inverter of Figure 12.13 requires fewer transistors overall (six), it contains four output driver devices that will consume considerable silicon area.

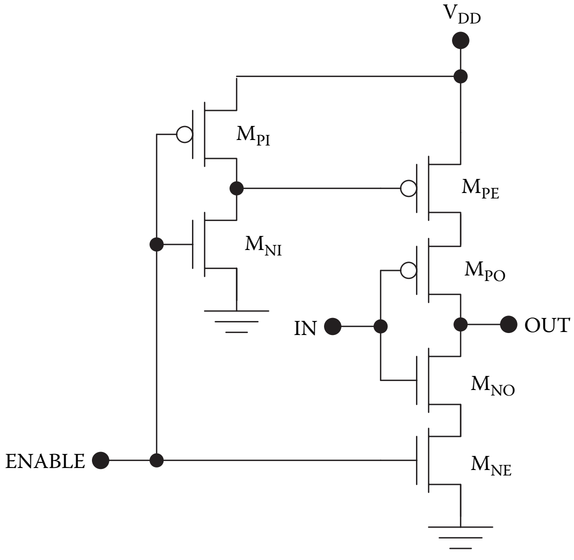


FIGURE 12.13
Tri-state CMOS inverter.

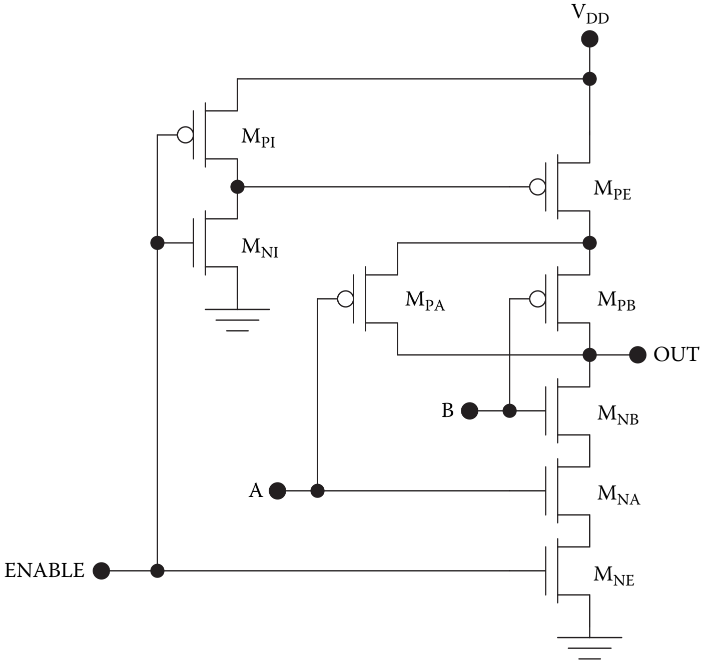


FIGURE 12.14
Tri-state CMOS NAND2 gate.

12.6 Interface Circuits

Digital systems usually combine different varieties of circuits operating at different voltage levels, and this requires the inclusion of interface circuits that translate, or shift, the voltage levels. At the present time, voltage

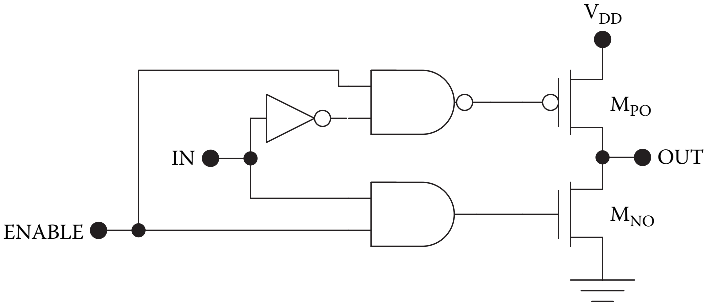


FIGURE 12.15
Tri-state CMOS driver.

translation circuits are most commonly used between CMOS circuits operating with different values of V_{DD}, and these will be described in more detail. Occasionally, however, it is even necessary to translate between CMOS and bipolar circuits or between gallium arsenide and silicon circuits.

12.6.1 High-Voltage CMOS to Low-Voltage CMOS

Many CMOS VLSI circuits operate with two or more supply voltages; this way output drivers can use a large V_{DD} for higher data rates and wider noise margins, whereas the core circuitry operates at a lower V_{DD} to reduce dissipation.

The translation from high-voltage CMOS (HV CMOS) to low-voltage CMOS (LV CMOS) can be accomplished using an asymmetric CMOS inverter as shown in Figure 12.16. The level translator circuit comprising M_{PT} and M_{NT} is connected to the lower supply voltage (V_{DDL}) but is designed to have a midpoint voltage (switching threshold) equal to one-half of the larger supply voltage (V_{DDH}). Thus,

$$V_M = \frac{V_{TN} + \left(V_{DDL} + V_{TP}\right)\sqrt{K_P / K_N}}{1 + \sqrt{K_P / K_N}} = \frac{V_{DDH}}{2}. \tag{12.14}$$

Rearranging, we find that the required ratio of device transconductance parameters is

$$K_R = \frac{K_N}{K_P} = \left(\frac{V_{DDL} - V_{DDH} / 2 - V_T}{V_{DDH} / 2 - V_T}\right)^2. \tag{12.15}$$

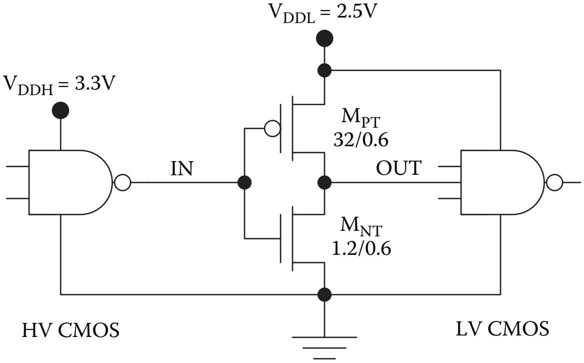


FIGURE 12.16
HV CMOS to LV CMOS level translator.

Of course it is not possible to match the switching threshold exactly to $V_{DDH}/2$ with one stage if $V_{DDH}/2 > V_{DDL} - V_T$.

Example 12.6. 3.3 to 2.5 V CMOS Level Translator

Design a CMOS level shifter to interface from 3.3 V circuits to 2.5 V circuits and plot the transfer characteristic for the circuit. Assume that $V_{TN} = |V_{TP}| = 0.5$ V for all circuits.

Solution: To interface from 3.3 V circuits to 2.5 V circuits, we can use an asymmetric CMOS inverter with

$$K_R = \frac{K_N}{K_P} = \left(\frac{V_{DDL} - V_{DDH}/2 - V_T}{V_{DDH}/2 - V_T}\right)^2 = \left(\frac{2.5V - 3.3V/2 - 0.5V}{3.3V/2 - 0.5V}\right)^2 = 0.093\ .$$

Therefore, the ratio of the transistor widths will be

$$\frac{W_N}{W_P} = \left(\frac{\mu_p}{\mu_n}\right)0.093 = \left(\frac{230cm^2/Vs}{580cm^2/Vs}\right)0.093 = 0.037 = \frac{1}{27}\ .$$

The voltage transfer characteristic can be calculated piecewise as follows:

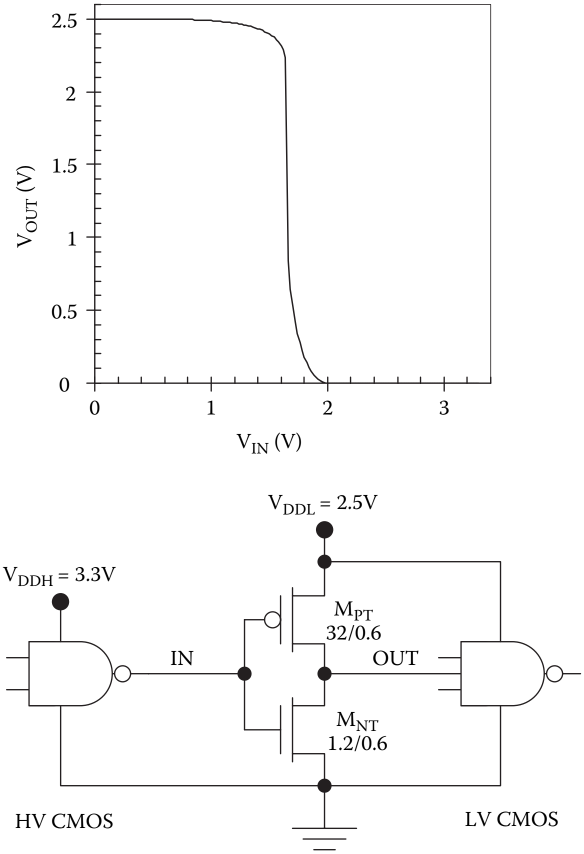
$$V_{OUT} = \begin{cases} 2.5V; & V_{IN} \le 0.5V \\ V_{IN} + 0.5V + \sqrt{\left(V_{IN} - 2.0V\right)^2 - \left(0.093\right)\left(V_{IN} - 0.5V\right)^2}; & 0.5V \le V_{IN} \le 1.65V \\ 1.25V; & V_{IN} = 1.65V \\ V_{IN} - 0.5V - \sqrt{\left(V_{IN} - 0.5V\right)^2 - \left(0.093\right)^{-1}\left(V_{IN} - 2.0V\right)^2}; & 1.65V \le V_{IN} \le 2.0V \\ 0; & V_{IN} \ge 2.0V \end{cases}$$

The translator circuit and its voltage transfer characteristic are displayed in Figure 12.17.

A disadvantage of using an asymmetric inverter for HV CMOS to LV CMOS translation is that the ratio of the device widths may be quite distant from unity, requiring a very large p-MOS transistor. Another approach, which avoids this difficulty, is to use a Schmitt trigger circuit as shown in Figure 12.18.

12.6.2 Low-Voltage CMOS to High-Voltage CMOS

Voltage translation from LV CMOS to HV CMOS may be accomplished using a voltage amplifier of the type illustrated in Figure 12.19. With $V_{IN} = 0$, M_{PO} and M_{NI} are linear, whereas M_{NO} and M_{PI} are cutoff, so $V_{OUT} = V_{DDH}$. With a logic "1" input, M_{NO} and M_{PI} are linear but M_{PO} and M_{NI} are cutoff so $V_{OUT} = 0$. Therefore, this translator is inverting like the circuits of the previous section.



All gate dimensions in µm

FIGURE 12.17
Example 3.3 to 2.5 V CMOS level translator.

12.7 SPICE Demonstrations

For the purpose of illustration, simulations were performed using Cadence Capture CIS 10.1.0 PSpice (Cadence Design Systems). The level 1 MOS transistor model parameters given in Tables 12.2 and 12.3 were used unless otherwise noted. The process transconductance parameters were calculated assuming an oxide thickness of 9 nm. For n-MOSFETS,

$$KP = \frac{(3.9)\left(8.85\times10^{-14}F/cm\right)\left(580cm^2V^{-1}s^{-1}\right)}{9\times10^{-7}cm} = 222\mu A/V^2, \quad (12.16)$$

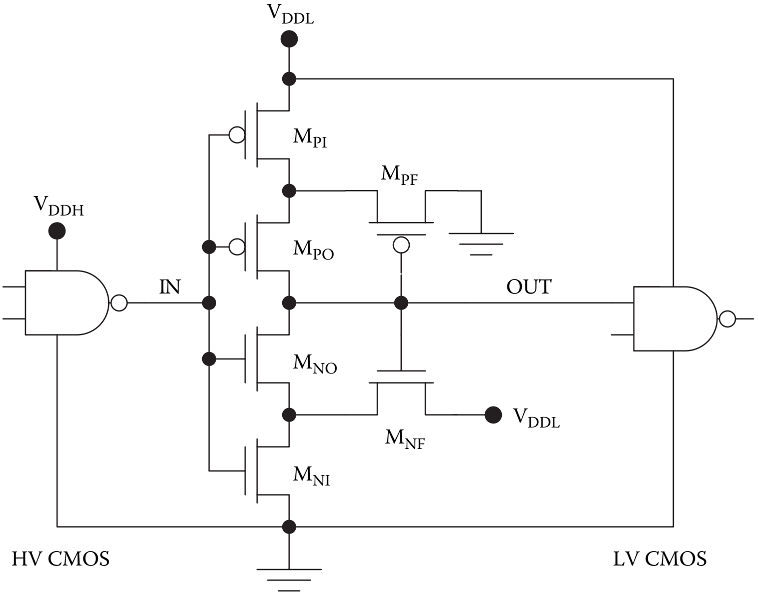


FIGURE 12.18
Schmitt trigger circuit used to connect between HV CMOS and LV CMOS with improved noise performance.

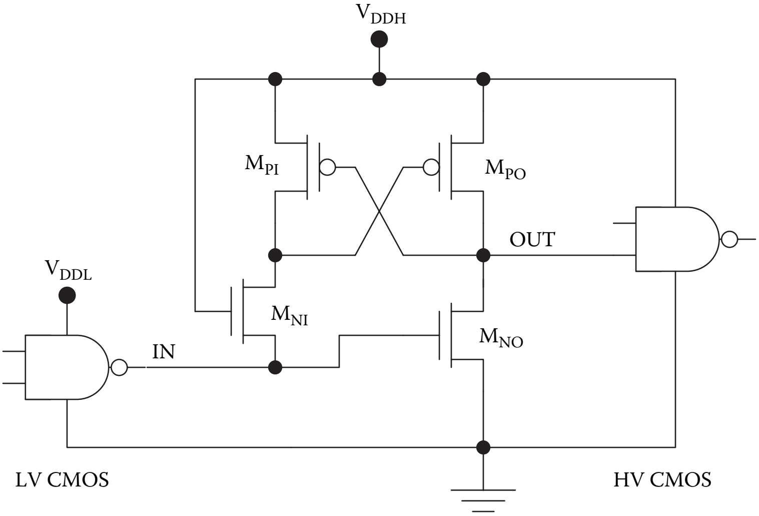


FIGURE 12.19
LV CMOS to HV CMOS level translator circuit.

TABLE 12.2

n-MOS Level 1 SPICE Parameters

Parameter	Value	Units
KP	222u	A/V^2
VTO	0.5	V
GAMMA	0.15	V$^{1/2}$
PHI	0.7	V
LAMBDA	0.05	
TOX	9n	m
NSUB	1E16	cm^{-3}
UO	580	cm^2/Vs
CGSO	0.38n	F/m
CGDO	0.38n	F/m

and for p-M-OSFETS,

$$KP = \frac{(3.9)\left(8.85\times10^{-14}F/cm\right)\left(230cm^2V^{-1}s^{-1}\right)}{9\times10^{-7}cm} = 88\mu A/V^2\,, \qquad (12.17)$$

The overlap capacitances per unit gate width were determined with the assumption that $L_{OV} = 0.1\mu m$:

$$\begin{aligned} CGSO &= \frac{(3.9)\left(8.85\times10^{-14}F/cm\right)\left(0.1\times10^{-4}cm\right)}{9\times10^{-7}cm} \\ &= 3.8pF/cm = 0.38nF/m \end{aligned} \qquad (12.18)$$

TABLE 12.3

p-MOS Level 1 SPICE Parameters

Parameter	Value	Units
KP	88u	A/V^2
VTO	–0.5	V
GAMMA	0.15	V$^{1/2}$
PHI	0.7	V
LAMBDA	0.05	
TOX	9n	m
NSUB	1E16	cm^{-3}
UO	230	cm^2/Vs
CGSO	0.38n	F/m
CGDO	0.38n	F/m

and

$$CGDO = \frac{(3.9)\left(8.85\times10^{-14}\,F/cm\right)\left(0.1\times10^{-4}\,cm\right)}{9\times10^{-7}\,cm}. \qquad (12.19)$$

$$= 3.8\,pF/cm = 0.38\,nF/m$$

The body effect coefficient was calculated from

$$GAMMA = \frac{\sqrt{2q\varepsilon_{Si}N_a}}{C_{ox}}$$

$$= \frac{\sqrt{2\left(1.602\times10^{-19}\,C\right)(11.9)\left(8.85\times10^{-14}\,F/cm\right)\left(10^{16}\,cm^{-3}\right)}}{(3.9)\left(8.85\times10^{-14}\,F/cm\right)/\,9\times10^{-7}\,cm}. \qquad (12.20)$$

$$\approx 0.15V^{1/2}$$

SPICE Example 12.1 CMOS Transmission Gate

The CMOS transmission gate of Figure 12.20 was investigated using a DC sweep of the current source IIN. With the enable voltage VENABLE set to 2.5 V, the input voltage is proportional to the input current as shown in Figure 12.21, and the "on" resistance of the transmission gate is 1.08 kΩ.

SPICE Example 12.2 Buffered CMOS

Transient simulations were performed to determine the overall propagation delay for three cases: an unbuffered CMOS inverter driving a 1 pF load as in Figure 12.22, a CMOS inverter with two buffer stages driving a 1 pF load as in Figure 12.23, and a CMOS inverter with four buffer stages driving a 1 pF load as in Figure 12.24. For the buffered cases, the transistor widths were sized up by a factor of three at each successive stage. The same abrupt input waveform was used in all three cases, with V1 = 0, V2 = 2.5 V, TD = 0, TR = 0, TF = 0, PW = 10 ns, and PER = 20 ns. The transient simulation results of Figure 12.25 show that the overall propagation delay is decreased significantly by the use of two scaled buffer stages. However, appending two more buffer stages gives little additional benefit in terms of the overall propagation delay with this load.

SPICE Example 12.3 Tri-State CMOS

The circuit of Figure 12.26 was used to study the behavior of a tri-state CMOS inverter. Pulse sources were used for the input and the enable voltages, and the input was set to twice times the frequency of the enable signal. The results of

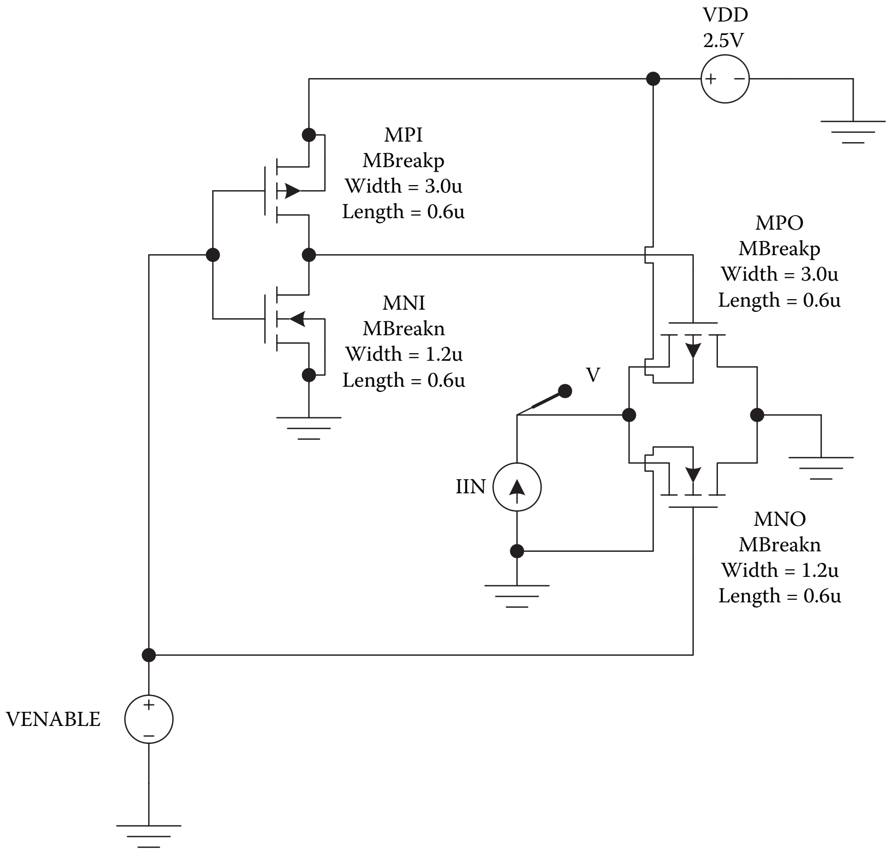


FIGURE 12.20
CMOS transmission gate for determination of the "on" resistance.

Figure 12.27 show that normal inverter action is obtained when the enable signal is high; otherwise, the inverter goes into the high impedance state and the output voltage is fixed at $V_{DD}/2$ by the resistor divider network.

12.8 Summary

Digital integrated circuits require the implementation of special input, output, and interface circuits as well as combinational logic gates, sequential logic gates, and memories. Inputs must have protection circuitry to prevent ESD damage during handling as well as transmission gates for enable/

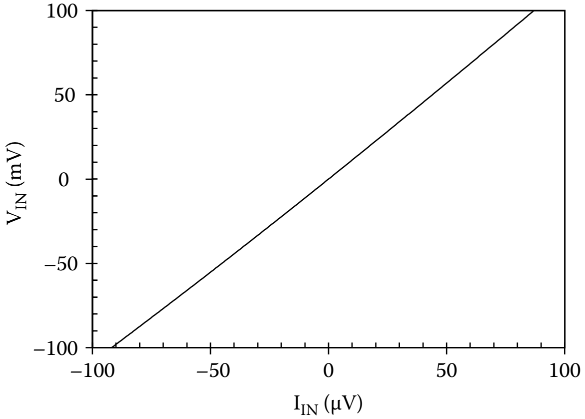


FIGURE 12.21
Input voltage as a function of input current for the enabled transmission gate of Figure 12.20.

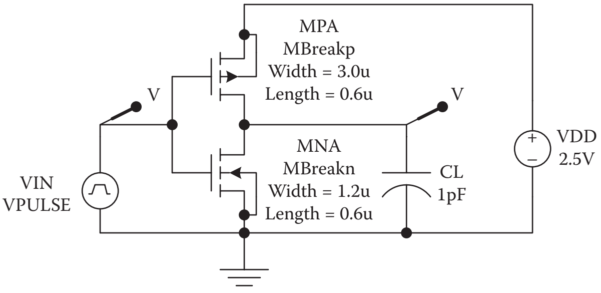


FIGURE 12.22
Unbuffered CMOS inverter driving a 1 pF load.

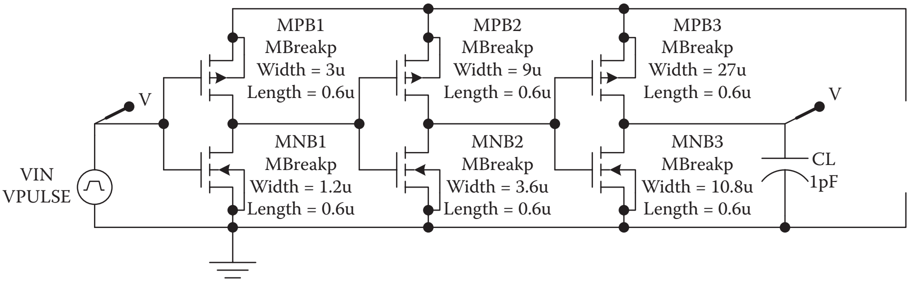


FIGURE 12.23
CMOS inverter driving a 1 pF load through two buffer stages.

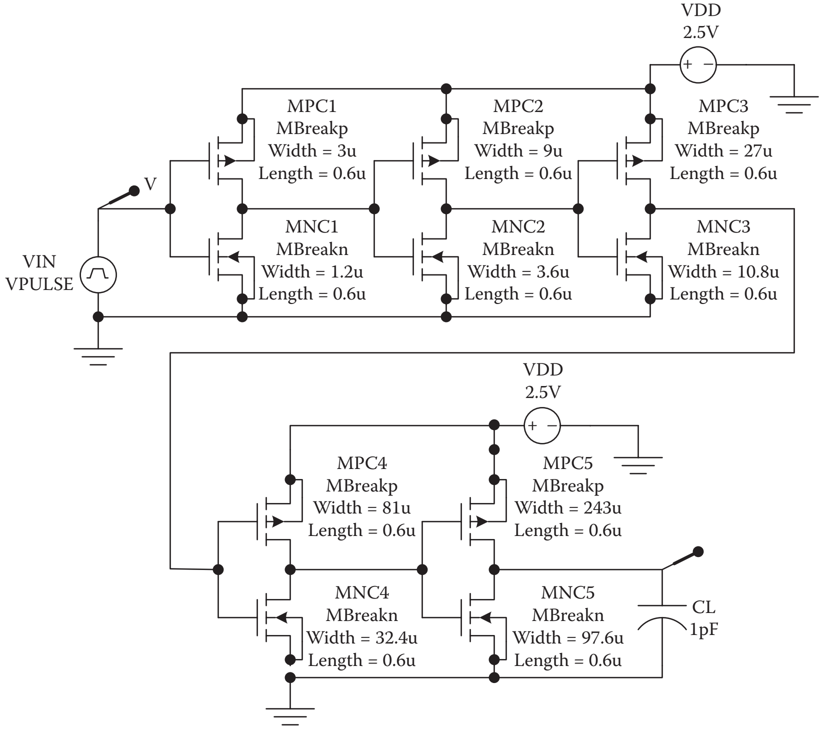


FIGURE 12.24
CMOS inverting driving a 1 pF load through four buffer stages.

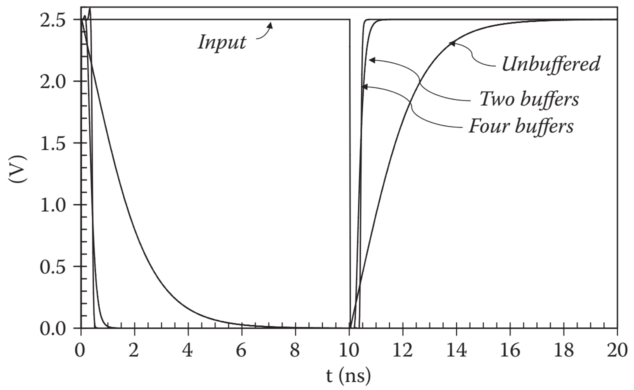


FIGURE 12.25
Transient simulation results for three cases: an unbuffered CMOS inverter driving a 1 pF load, a CMOS inverter with two buffer stages driving a 1 pF load, and a CMOS inverter with four buffer stages driving a 1 pF load.

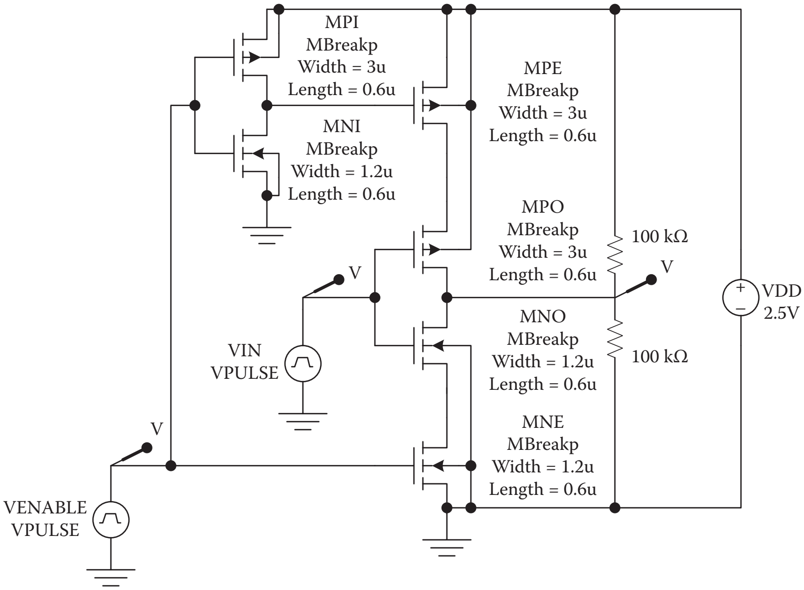


FIGURE 12.26
Tri-state CMOS inverter.

disable operation. Output pins generally require high-current output drivers and tri-state operation to allow compatibility with busses. Interface circuits are needed for voltage level shifting between circuits operating with different voltages.

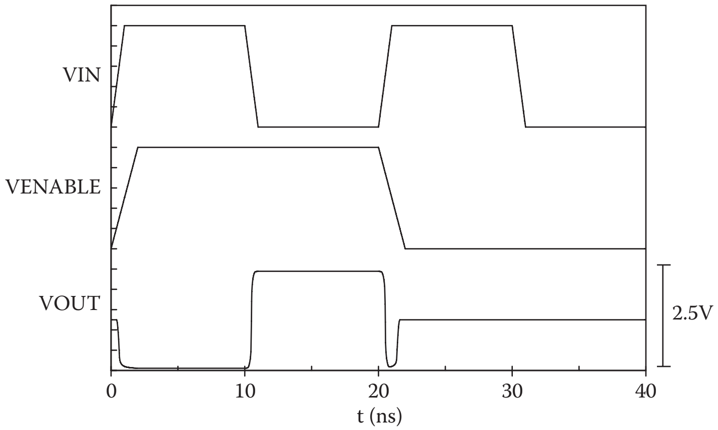


FIGURE 12.27
Transient response for the tri-state CMOS inverter of Figure 12.26.

12.9 Practical Perspective

For practical perspective articles, see the dynamic website at http://www.engr.uconn.edu/ece/books/ayers.

12.10 Exercises

E12.1. Determine the "on" resistance as a function of V_{IN} for the symmetric CMOS transmission gate of Figure 12.28 with $V_{ENABLE} = 2.5V$.

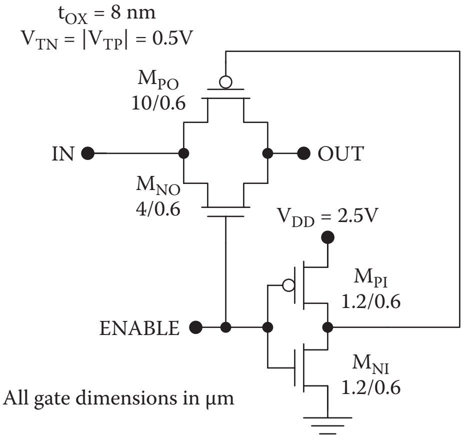


FIGURE 12.28
CMOS transmission gate for analysis of the "on" resistance (see Exercise E12.1).

E12.2. Determine the "on" resistance as a function of V_{IN} for the asymmetric CMOS transmission gate of Figure 12.29 with $V_{ENABLE} = 2.5V$.

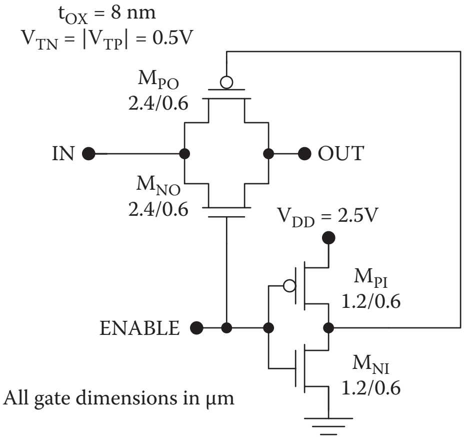


FIGURE 12.29
CMOS transmission gate (see Exercise E12.2).

E12.3. Choose the widths of the devices M_{PO} and M_{NO} in Figure 12.30 so the maximum "on" resistance is $1k\Omega$ with $V_{ENABLE} = 1.5V$.

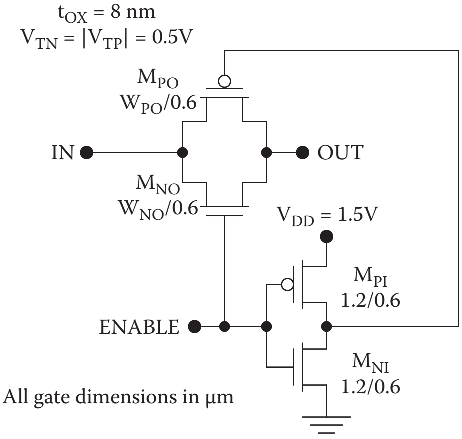


FIGURE 12.30
Transmission gate for the design of the transistor widths (see Exercise E12.3).

E12.4. The inverter shown in Figure 12.31 drives a 5 pF load through three buffer stages, with a scale factor of four at each stage. The input to the inverter on the left makes abrupt voltage transitions. (1) Find the propagation delay for the inverter driving the 5 pF load directly, without buffering. (2) Find the input capacitances for each of the three buffer stages. (3) Find the four propagation delays and the total delay for the inverter with the three buffers.

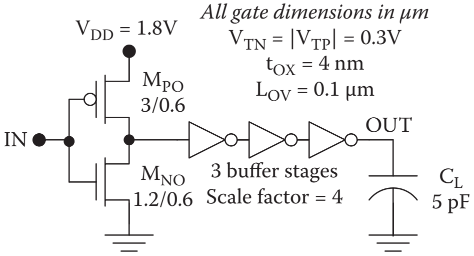


FIGURE 12.31
A inverter with three buffer stages driving a capacitive load (see Exercise E12.4).

E12.5. The inverter of Figure 12.32 drives an off-chip load through three buffers as shown. Using the short-channel MOSFET equations, find the propagation delays for each of the four inverters and the total propagation delay.

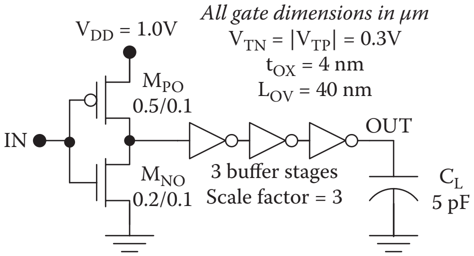


FIGURE 12.32
Buffer stages driving a 5 pF load (see Exercise E12.5).

E12.6. An inverter drives a 1 pF load through buffer stages having a uniform scale factor of five as shown in Figure 12.33. How many buffer stages are needed so that $t_{PLH} \leq 600ps$ and $t_{PHL} \leq 600ps$?

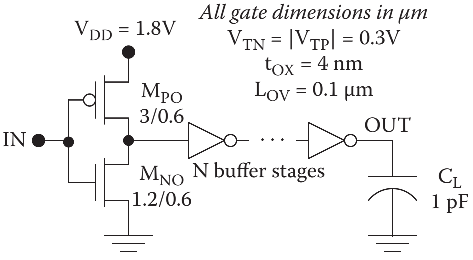


FIGURE 12.33
Inverter with N buffer stages to drive a 1 pF load (see Exercise E12.6).

E12.7. For the arrangement shown in Figure 12.34, determine and plot the total propagation delay as a function of the number of buffer stages. The buffer stages use a uniform scale factor of five. What is the ideal number of buffers in terms of delay?

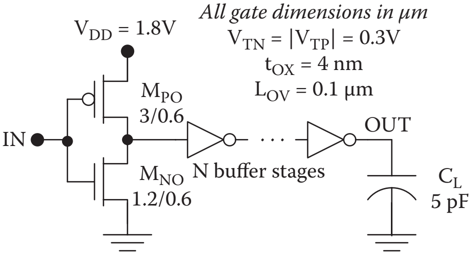


FIGURE 12.34
CMOS inverter with N buffer stages to drive a load capacitance (see Exercise E12.7).

E12.8. For the buffered configuration of Figure 12.35, (1) determine and plot the total propagation delay as a function of the number of buffer stages, assuming a uniform scale factor of five, and (2) determine and plot the product of the circuit area and the total propagation delay as a function of the number of buffer stages. Use the short-channel MOSFET equations.

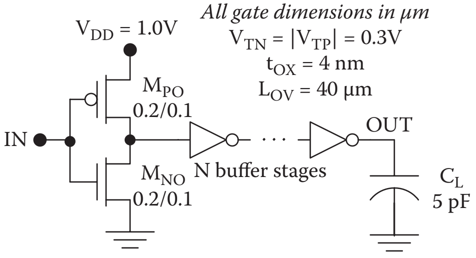


FIGURE 12.35
Inverter with buffers (see Exercise E12.8).

E12.9. Design a level translation circuit to go from 3.3 V CMOS to 1.0 V CMOS, using one or more asymmetric inverters. Calculate and plot the overall voltage transfer characteristic for your level translator. $V_{TN} = |V_{TP}| = 0.3V$ and $L = 0.5\mu m$ for all devices. You may use any or all of the following supply voltages: 1.0, 1.2, 1.5, 1.8, 2.5, and 3.3 V.

For additional exercise problems, see the dynamic website at http://www.engr.uconn.edu/ece/books/ayers.

References

1. Avery, L.R., ESD protection structure issues and design for custom integrated circuits. *Proc. IEEE Custom Integrated Circuits Conf.*, 1–27, 1988.
2. Ker, M.-D., and Chuang, C.-H., ESD protection circuits with novel MOS-bounded diode structures. *IEEE Int. Symp. Circuits Syst.*, 533–536, 2002.
3. Ker, M.-D., and Jiang, H.-C., Whole-chip ESD protection strategy for CMOS integrated circuits in nanotechnology. *Proc. 1st IEEE Conf. Nanotechnol.*, 325–330, 2001.
4. Hew, C.F., and Thorbjornsen, A.R., A macromodel for a CMOS transmission gate. *Proc. 8th University/Government/Industry Microlectronics Symp.*, 178–181, 1989.
5. Cherkauer, B.S. and Friedman, E.G., Design of tapered buffers with local interconnect capacitance. *IEEE J. Solid State Circuits*, 30, 151–155, 1995.

Appendix A

List of Symbols

Symbol	Description	Units
α	Switching activity factor	
α_n	Switching activity factor for the nth node	
ΔL	Reduction in channel length	cm
ε	Permittivity	F/cm
ε_0	Permittivity of free space	F/cm
ε_{DI}	Permittivity of the dielectric	F/cm
ε_{ox}	Permittivity of SiO_2	F/cm
ε_s	Permittivity of semiconductor	F/cm
ε_{Si}	Permittivity of silicon	F/cm
ϕ_{MS}	Metal-semiconductor work function difference	V
γ	Body effect coefficient	$V^{1/2}$
Γ_N	n-MOS delay parameter	V^{-1}
Γ_{Neff}	Short-channel n-MOS effective delay parameter	V^{-1}
Γ_P	p-MOS delay parameter	V^{-1}
Γ_{Peff}	Short-channel p-MOS effective delay parameter	V^{-1}
κ	Dielectric constant	
κ	MOS transistor scale factor	
λ	Optical wavelength	cm
λ	channel length modulation parameter	V^{-1}
λ_N	n-MOS channel length modulation parameter	V^{-1}
λ_P	p-MOS channel length modulation parameter	V^{-1}
μ	Mobility (electron or hole)	cm^2/Vs
μ_{DI}	Permeability of the dielectric	H/cm
μ_n	Electron mobility	cm^2/Vs
μ_p	Hole mobility	cm^2/Vs
ρ	Resistivity	Ω
ρ	Space charge density	C/cm^2
σ	Conductivity	A/V
τ_F	Effective forward lifetime	
τ_n	Electron lifetime	s
τ_p	Hole lifetime	s
τ_{SC}	Space charge lifetime	s

(continued)

(continued)

Symbol	Description	Units
ϕ_{ms}	Metal-semiconductor work function difference	V
ψ_i	Electric potential	V
a	Lattice constant	cm
A	Chip area	cm^2
A	Junction area	cm^2
c	Capacitance per unit length	F/cm
C	Capacitance	F
C	Control input	
C_{bot}	Source, drain junction capacitance per unit area on bottom	F/cm^2
C_{bot0}	Source, drain zero-bias bottom depletion capacitance	
C_D	Capacitance between drain and ground	F
C_{dm}	Maximum depletion layer capacitance per unit area	F/cm^2
C_G	Capacitance between gate and ground	F
C_{gb}	Gate-body capacitance	F
C_{gd}	Gate-drain capacitance	F
C_{gs}	Gate-source capacitance	F
C_{in}	Input capacitance	F
$C_{interconnect}$	Interconnect capacitance	F
C_{J0}	Zero-bias depletion capacitance	F
C_L	Load capacitance	F
$C_{L,max}$	Maximum load capacitance	F
C_{Mi}	Miller input capacitance	F
C_{Mo}	Miller output capacitance	F
C_{out}	Output capacitance	C
C_{ox}	Oxide capacitance per unit area	F/cm^2
C_{sw}	Source, drain depletion capacitance per unit area on sidewall	F/cm^2
C_{sw0}	Source, drain zero bias sidewall depletion capacitance	F/cm^2
C_T	Depletion layer capacitance (transition capacitance)	F
C_Y	Soft node capacitance	C
d	Depth of focus	cm
D_0	Areal density of defects	cm^{-2}
D_n	Electron diffusivity	cm^2/Vs
D_p	Hole diffusivity	cm^2/Vs
E	Electric field intensity	V/cm
E_a	Activation energy	eV
E_c	Edge of the conduction band	eV
E_f	Fermi level	eV
E_g	Energy gap	eV
E_i	Intrinsic Fermi level	eV
E_v	Edge of the valence band	eV
f	Frequency, switching frequency	s^{-1}

(continued)

Symbol	Description	Units
f_{CLK}	Clock frequency	s^{-1}
f_M	Ring oscillator frequency with M inverters	s^{-1}
G	Generation rate for electron-hole pairs	$cm^{-3}s^{-1}$
i_n	Electron current	A
I_D	Drain current	A
I_{DD}	Drain supply current	A
I_{DDH}	Drain supply current with the output high	A
I_{DDL}	Drain supply current with the output low	A
I_{DN}	n-MOS drain current	A
I_{DP}	p-MOS drain current	A
I_{DSATn}	n-MOS saturated drain current	A
I_{DSATp}	p-MOS saturated drain current	A
I_F	Forward current	A
I_{gen}	Generation current	A
I_{leak}	Leakage current	A
I_{PN}	p-n junction leakage current	A
I_{OH}	Output high current	A
I_{OL}	Output low current	A
I_{pn}	p-n junction leakage current	A
I_S	Reverse saturation current	A
$I_{Subthreshold}$	Subthreshold current	A
J	Current density	A/cm^2
J	Energy	J
J_n	Electron current density	A/cm^2
J_p	Hole current density	A/cm^2
J_{switch}	Switching energy	J
k	Boltzmann constant	J/K
k'	Process transconductance parameter	A/V^2
k_N'	n-MOS process transconductance parameter	A/V^2
k_P'	p-MOS process transconductance parameter	A/V^2
K	Device transconductance parameter	A/V^2
K_N	n-MOS device transconductance parameter	A/V^2
K_P	p-MOS device transconductance parameter	A/V^2
K_R	Ratio of device transconductance parameters	
l	Inductance per unit length	H/cm
L	Gate length	cm
L′	Reduced channel length	cm
L	Inductance	H
L_D	Length of drain	cm
L_n	Electron diffusion length	cm
L_N	n-MOS gate length	cm
L_{OV}	Overlap of gate with source, drain	cm

(continued)

(continued)

Symbol	Description	Units
L_p	Hole diffusion length	cm
L_P	p-MOS gate length	cm
L_S	Length of source	cm
LS	Logic swing	V
m	Subthreshold parameter (capacitance ratio parameter)	
m	Grading coefficient	
M	Fan-in	
M	Number of inverters in a ring oscillator	
n	Electron concentration	cm^{-3}
n	Empirical velocity saturation coefficient (electrons or holes)	
$\bar{n}$	Equilibrium electron concentration	cm^{-3}
n'	Excess electron concentration	cm^{-3}
n_p	Electron concentration in p-type material	cm^{-3}
n_p'	Excess electron concentration in p-type material	cm^{-3}
n_i	Intrinsic carrier concentration	cm^{-3}
N	Fan-out	
N	Emission coefficient (p-n junction)	
N	Number of die on the wafer	
N_a	Acceptor concentration	cm^{-3}
N_d	Donor concentration	cm^{-3}
N_C	Effective density of states at the edge of the conduction band	cm^{-3}
N_D	Number of defects on the wafer	
N_G	Number of good die on the wafer	
N_{MAX}	Maximum fan-out	
N_S	Number of logic stages	
N_V	Effective density of states at the edge of the valence band	cm^{-3}
NA	Numerical aperture	
p	Hole concentration	cm^{-3}
$\bar{p}$	Equilibrium hole concentration	cm^{-3}
p'	Excess hole concentration	cm^{-3}
P_{AC}	Dynamic dissipation	W
P_{DC}	Static dissipation	W
P_L	Output low-power dissipation	W
P_H	Output high-power dissipation	W
Ppn	p-n junction leakage power	W
P_{sc}	Short-circuit power	W
P_{switch}	Capacitance switching power	W
$P_{subthreshold}$	Subthreshold power	W
PDP	Power-delay product	J
q	Electronic charge	C
$q\chi$	Semiconductor electron affinity	eV

(continued)

Symbol	Description	Units
$q\phi_F$	Bulk difference between intrinsic and actual Fermi level	eV
$q\phi_m$	Metal work function	eV
$q\phi_s$	Semiconductor work function	eV
$q\Psi_s$	Band bending at the surface	eV
Q_B	Semiconductor depletion charge per unit area	C/cm^2
Q_B	Stored charge in the base	C
Q_i	Inversion layer charge	C/cm^2
Q_{II}	Ion-implanted charge	Ccm^{-2}
Q_{ox}	Oxide charge	C/cm^2
r	Resistance per unit length	Ω/cm
R	Resistance	Ω
R	Recombination rate for minority carriers	$cm^{-3}s^{-1}$
R_{ON}	“on” resistance	Ω
R_{ONN}	n-MOS drain “on” resistance	Ω
R_{ONP}	p-MOS drain “on” resistance	Ω
R_L	Load resistance	Ω
S	Subthreshold swing	V
s	Scaling factor	
t	Time	s
T	Temperature	K
T	Period of ring oscillator	s
t_0	Logic “0” transfer time	s
t_1	Logic “1” transfer time	s
t_F	Fall time	s
t_F'	Extrinsic fall time	s
t_{F1}	Component of fall time with n-MOS saturated	s
t_{F2}	Component of fall time with n-MOS linear	s
t_{FI}	Fall time of input signal	s
t_{ox}	Oxide thickness	cm
t_P	Average propagation delay	s
$t_{P,max}$	Maximum propagation delay	s
t_{PHL}	High-to-low propagation delay	s
t'_{PHL}	Extrinsic high-to-low propagation delay	s
t_{PHL1}	Component of t_{PHL} with n-MOS saturated	s
t_{PHL2}	Component of t_{PHL} with n-MOS linear	s
$t_{PHL,int}$	Intrinsic high-to-low propagation delay	s
t_{PLH}	Low-to-high propagation delay	s
t'_{PLH}	Extrinsic low-to-high propagation delay	s
t_{PLH1}	Component of t_{PLH} with p-MOS saturated	s
t_{PLH2}	Component of t_{PLH} with p-MOS linear	s
$t_{PLH,int}$	Intrinsic low-to-high propagation delay	s
t_R	Rise time	s
t_R'	Extrinsic rise time	s

(continued)

(continued)

Symbol	Description	Units
t_{R1}	Component of rise time with p-MOS saturated	s
t_{R2}	Component of rise time with p-MOS linear	s
t_{ret}	Retention time	s
t_{RI}	Rise time of input signal	s
t_s	Storage delay time	s
t_t	Transit time	s
v	Carrier velocity (electron or hole)	cm/s
V	Electric potential	V
V_{bi}	Built-in voltage	V
V_B	Breakdown voltage	V
V_{BS}	Body-to-source bias voltage	V
V_D	Diode turn-on voltage	V
V_{DD}	Drain supply voltage	V
V_{DS}	Drain-to-source voltage	V
V_{DSAT}	Saturation drain-to-source voltage	V
V_F	Forward voltage	V
V_{GS}	Gate-to-source voltage	V
V_{IH}	Input high voltage	V
V_{IL}	Input low voltage	V
V_{IN}	Input voltage	V
V_L	Lower trip voltage	V
V_M	Midpoint voltage or switching threshold	V
V_{NMH}	High noise margin	V
V_{NML}	Low noise margin	V
V_{OH}	Output high voltage	V
V_{OL}	Output low voltage	V
V_{OUT}	Output voltage	V
V_R	Reverse voltage	V
v_{sat}	Saturation velocity	cm/s
v_{satn}	Electron saturation velocity	cm/s
v_{satp}	Hole saturation velocity	cm/s
V_{ss}	Steady-state voltage	V
V_{SS}	Source supply	V
V_T	Threshold voltage	V
V_{TH}	V_T for high threshold devices	V
V_{TL}	Threshold voltage for pull-up device	V
V_{TL}	V_T for low-threshold devices	V
V_{TN}	n-MOS threshold voltage	V
V_{TO}	Zero-bias threshold voltage	V
V_{TO}	Threshold voltage for pull-down device	V
V_{TP}	p-MOS threshold voltage	V
V_U	Upper trip voltage	V

(continued)

Symbol	Description	Units
W	Depletion width	cm
W	Gate width	cm
W_B	Base width	cm
W_D	Drain depletion width	cm
W_{dm}	Depletion width in silicon under inversion	cm
W_N	n-MOS gate width	cm
W_P	p-MOS gate width	cm
W_S	Source depletion width	cm
X	One-half the minimum feature size 2X	µm
x_j	Junction depth	cm
x_n	Depletion width in the n-type side of a p-n junction	cm
x_p	Depletion width on the p-type side of a p-n junction	cm
Y	Yield	
Y_a	Assembly yield	
Y_{bi}	Burn-in yield	
Y_p	Process yield	
Y_w	Wafer yield	

Appendix B

International System of Units

Quantity	Units	Symbol	Equivalent units
Length	Meter	M	
Mass	Kilogram	Kg	
Time	Second	s	
Temperature	Kelvin	K	
Charge	Coulomb	C	
Current	Ampere	A	C/s
Potential	Volt	V	J/C
Power	Watt	W	J/s
Energy*	Joule	J	Nm
Conductance	Siemen	S	A/V
Resistance	Ohm	Ω	V/A
Capacitance	Farad	F	C/V
Magnetic flux	Weber	Wb	Vs
Flux density	Tesla	T	Wb/m^2
Inductance	Henry	H	Wb/A
Frequency	Hertz	Hz	1/s
Force	Newton	N	kgm/s^2
Pressure	Pascal	Pa	N/m^2

* In work with semiconductors and devices, it is common to use electron-volt (eV) units for energy. $1eV = 1.602 \times 10^{-19} J$.

Appendix C

Unit Prefixes

Multiplier	Prefix	Symbol
10^{18}	Exa	E
10^{15}	Peta	P
10^{12}	Tera	T
10^{9}	Giga	G
10^{6}	Mega	M
10^{3}	Kilo	k
10^{2}	Hecto	h
10^{1}	Deka	da
10^{-1}	Deci	d
10^{-2}	Centi	c
10^{-3}	Milli	m
10^{-6}	Micro	μ
10^{-9}	Nano	n
10^{-12}	Pico	p
10^{-15}	Femto	f
10^{-18}	Atto	A

Appendix D

Greek Alphabet

Letter	Lowercase	Uppercase
Alpha	α	Α
Beta	β	Β
Gamma	γ	Γ
Delta	δ	Δ
Epsilon	ε	Ε
Zeta	ζ	Ζ
Eta	η	Η
Theta	θ	Θ
Iota	ι	Ι
Kappa	κ	Κ
Lambda	λ	Λ
Mu	μ	Μ
Nu	ν	Ν
Xi	ξ	Ξ
Omicron	ο	Ο
Pi	π	Π
Rho	ρ	Ρ
Sigma	σ	Σ
Tau	τ	Τ
Upsilon	υ	Υ
Phi	ϕ	Φ
Chi	χ	Χ
Psi	ψ	Ψ
Omega	ω	Ω

Appendix E

Physical Constants

Quantity	Symbol	Value
Avogadro constant	N_A	$6.022 \times 10^{23}\, mol^{-1}$
Bohr radius	a_B	$5.292 \times 10^{-11}\, m$
Boltzmann constant*	k	$1.381 \times 10^{-23}\, J / K$
Electron rest mass	m_0	$9.110 \times 10^{-31}\, kg$
Electron charge	q	$1.602 \times 10^{-19}\, C$
Gas constant	R	$1.987\, cal / molK$
Permeability of free space	μ_0	$4\pi \times 10^{-9}\, H / cm$
Permittivity of free space	ε_0	$8.85 \times 10^{-14}\, F / cm$
Planck constant	h	$6.626 \times 10^{-34}\, Js$
Reduced Planck constant	η	$1.0546 \times 10^{-34}\, Js$
Speed of light in free space	c	$3.00 \times 10^{10}\, cm / s$

* In work with semiconductors, the value $k = 8.62 \times 10^{-5}\, eV / K$ is often used.

Appendix F

Properties of Si and Ge at 300 K

Property	**Si**	**Ge**
Atomic density (cm^{-3})	5.0×10^{22}	4.42×10^{22}
Density (g/cm^3)	2.329	5.327
Dielectric constant, ε_s	11.9	16.0
Effective density of states in the conduction band, N_C (cm^{-3})	2.8×10^{19}	1.04×10^{19}
Effective density of states in the valence band, N_V (cm^{-3})	1.04×10^{19}	6.0×10^{18}
Electron affinity χ (V)	4.05	4.00
Energy gap, E_g (eV)	1.12	0.66
Index of refraction, n	3.42	4.0
Intrinsic carrier concentration, n_i (cm^{-3})	1.45×10^{10}	2.4×10^{13}
Lattice constant, a (nm)	0.543108	0.564613
Mobility, bulk ($cm^2V^{-1}s^{-1}$)	$\mu_n = 1500$ $\mu_p = 450$	$\mu_n = 3900$ $\mu_p = 1900$
Mobility, MOSFET ($cm^2V^{-1}s^{-1}$)	$\mu_n = 580$ $\mu_p = 230$	
Specific heat (J/g°C)	0.713	0.31
Thermal conductivity ($Wcm^{-1}/°C$)	1.56	0.6

Appendix G

Properties of SiO_2 at 300 K

Property	**SiO_2**
Density (g/cm^3)	2.27
Dielectric constant, ε_r	3.9
Dielectric strength (V/cm)	10^7
Electron affinity χ (V)	0.9
Energy gap, E_g (eV)	9
Index of refraction, n	1.46
Specific heat (J/g^{o}C)	1.0
Thermal conductivity (W/cm^{o}C)	1.4×10^{-2}

Appendix H

Important Equations

n-MOS Transistor

$$V_{TO} = \phi_{MS} - 2\phi_F - \frac{Q_B}{C_{ox}} - \frac{Q_{ox}}{C_{ox}} - \frac{Q_{II}}{C_{ox}} \tag{H.1}$$

$$V_T = V_{TO} + \gamma\left(\sqrt{\left|V_{BS} + 2\phi_F\right|} - \sqrt{\left|2\phi_F\right|}\right) \tag{H.2}$$

$$K_N = k'_N \frac{W_N}{L_N} = \frac{\mu_n \varepsilon_{ox}}{t_{ox}} \frac{W_N}{L_N} \tag{H.3}$$

$$\begin{gathered} I_D = K_N\left[\left(V_{GS} - V_T\right)V_{DS} - \frac{V_{DS}^2}{2}\right]\left(1 + \lambda V_{DS}\right); \\ \left(V_{GS} \geq V_{TN}\right) \mathit{and} \left(V_{GS} - V_{TN}\right) \geq V_{DS} \end{gathered} \tag{H.4}$$

$$I_D = \frac{K_N}{2}\left(V_{GS} - V_{TN}\right)^2\left(1 + \lambda V_{DS}\right); \ \left(V_{GS} \geq V_{TN}\right) \mathit{and} \left(V_{GS} - V_{TN}\right) \leq V_{DS} \tag{H.5}$$

$$I_D \approx K\left(m - 1\right)\left(\frac{kT}{q}\right)^2 \exp\left(\frac{q\left(V_{GS} - V_T\right)}{mkT}\right) \mathit{(subthreshold)} \tag{H.6}$$

Short-Channel n-MOS Transistor

$$I_D = \frac{\mu_n C_{ox}\left(W / L\right)\left[\left(V_{GS} - V_{TN}\right)V_{DS} - V_{DS}^2 / 2\right]}{1 + \left(\mu_n V_{DS} / v_{satn} L\right)} \text{ (linear)} \tag{H.7}$$

$$V_{DSAT} = \frac{2\left(V_{GS} - V_{TN}\right)}{1 + \sqrt{1 + 2\mu_n\left(V_{GS} - V_{TN}\right)/\left(v_{satn}L\right)}} \tag{H.8}$$

$$I_D = C_{ox}Wv_{satn}\left(V_{GS} - V_{TN}\right)\frac{\sqrt{1 + 2\mu_n\left(V_{GS} - V_{TN}\right)/\left(v_{satn}L\right)} - 1}{\sqrt{1 + 2\mu_n\left(V_{GS} - V_{TN}\right)/\left(v_{satn}L\right)} + 1} \quad \text{(saturation)} \tag{H.9}$$

CMOS

$$V_M = \frac{V_{TN} + \left(V_{DD} + V_{TP}\right)\sqrt{K_P / K_N}}{1 + \sqrt{K_P / K_N}} \tag{H.10}$$

$$t_{PHL} = \frac{C_L}{K_N}\Gamma_N = \frac{C_L}{K_N\left(V_{DD} - V_{TN}\right)}\left[\frac{2V_{TN}}{\left(V_{DD} - V_{TN}\right)} + \ln\left(\frac{3V_{DD} - 4V_{TN}}{V_{DD}}\right)\right] \tag{H.11}$$

$$t_{PLH} = \frac{C_L}{K_P}\Gamma_P = \frac{C_L}{K_P\left(V_{DD} + V_{TP}\right)}\left[\frac{-2V_{TP}}{\left(V_{DD} + V_{TP}\right)} + \ln\left(\frac{3V_{DD} + 4V_{TP}}{V_{DD}}\right)\right] \tag{H.12}$$

$$P_{switch} = \alpha f_{CLK}C_L V_{DD}^2 \tag{H.13}$$

Short-Channel CMOS

$$t_{PHL} \approx \frac{C_L V_{DD}}{2C_{ox}W_N v_{satn}\left(V_{DD} - V_{TN}\right)}\frac{\sqrt{1 + 2\mu_n\left(V_{DD} - V_{TN}\right)/\left(v_{satn}L_N\right)} + 1}{\sqrt{1 + 2\mu_n\left(V_{DD} - V_{TN}\right)/\left(v_{satn}L_N\right)} - 1} \tag{H.14}$$

$$t_{PLH} \approx \frac{C_L V_{DD}}{2C_{ox}W_P v_{satp}\left(V_{DD} + V_{TP}\right)}\frac{\sqrt{1 + 2\mu_p\left(V_{DD} - V_{TP}\right)/\left(v_{satp}L_P\right)} + 1}{\sqrt{1 + 2\mu_p\left(V_{DD} - V_{TP}\right)/\left(v_{satp}L_P\right)} - 1} \tag{H.15}$$

Appendix I

Design Rules

The physical design of transistors, wires, and other integrated circuit components is governed by a set of design rules. These design rules, which dictate the dimensions, physical size, and device capacitances, are of three types: minimum dimensions, minimum spacings, and minimum surrounds. These minimum values are inextricably tied to the fabrication process as well as the lithographic process used for pattern transfer.

Design rules may be scalable or absolute. Scalable rules are stated in terms of X (where the minimum feature size is 2X)*, whereas absolute design rules are stated in units of length (nanometers). Scalable rules have the advantage that they can be applied to different process lines having different values of X, but they may not be simultaneously optimized for different values of X. Some design rules do not scale with X, so worst-case values must be used to produce a scalable rule set. In practice, both scalable and absolute design rules are in use today. Examples of scalable design rules sets are those used by the VLSI prototyping service MOSIS [1].

Generally speaking, there are three classes of design rules: (1) minimum widths, (2) minimum spacings, and (3) minimum surrounds. Table I.1 lists the most important design rules of these types, along with a particular set of values adapted from the MOSIS rule set and used for the examples and exercises in this book. Advanced design work will always be based on similar rules, possibly with the inclusion of additional rules, but the numerical values (either scalable, in terms of X, or absolute, in terms of nanometers) will tend to be different.

These design rules are illustrated in the following series of figures with an orientation to the basic n-well CMOS process described previously. The basic layers and layout legends for such a scalable n-well CMOS process are summarized in Table I.2.

Note that the active layer defines the placement of silicon nitride, which in turn is used to pattern shallow trench oxide; the shallow trench oxide is grown wherever the nitride is *absent*. Therefore, channel regions are defined by the overlap of the active and polysilicon layers.

* 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.

TABLE I.1

Layout Design Rules along with Values Adopted for This Book

Rule	Description	Value
Minimum dimensions		
L1	Gate length/polysilicon width	2X
L2	Extension of polysilicon gate beyond active region	1X
L3	Width of contact window	2X
L4	Width of active region	3X
L5	Width of implanted region	3X
L6	Width of metal 1	3X
L7	Width of metal 2	3X
Minimum separations		
D1	S pacing between polysilicon gates/ wires	2X
D2	Spacing between polysilicon gate and S/D contact window	2X
D3	Spacing between contacts	2X
D4	Spacing between active regions	3X
D5	Spacing between implanted regions of same type	3X
D6	Spacing between metal 1 wires	3X
D7	Spacing between metal 2 wires	4X
D8	Spacing between implanted regions of opposite type	5X
Minimum surrounds		
S1	Active region surrounding contact window	1X
S2	Metal 1 surrounding contact window	1X
S3	Metal 2 surrounding contact window	1X
S4	Polysilicon surrounding contact window	1X
S4	nselect or pselect surrounding contact window	1X
S6	nselect or pselect surrounding active region	1X
S7	n-Well surrounding p-MOS active region	5X

A single mask is used to pattern the polysilicon wires, although these wires exist with both p-type doping and n-type doping because polysilicon is doped simultaneously with the source and drain regions of the MOSFETs, as required by the *self-aligned process*. However, metal or silicide straps are generally placed over the polysilicon to alleviate problems associated with the transition from p-type polysilicon to n-type polysilicon.

Often simplified layouts will be used for the purpose of illustration, and these use the simplified legend of Table I.3.

The minimum line width 2X is the smallest dimension permitted for any feature in the layout. 2X is also called the *minimum feature size*. At the time of this writing, a minimum feature size of 45 nm is used in production, and we say that we are at the *45 nm technology node*. Technologically, the minimum feature size corresponds to the minimum width for a polysilicon line. For example, with 0.1 μm technology, the minimum polysilicon line

TABLE I.2

Detailed Layout Legend

Physical layer	Name	Layout symbol
n-well	n-well	□
Silicon nitride	Active	▧
Polysilicon	Poly 1	▩
p+ implant	p select	░
n+ implant	n select	▓
Contact cut	Contact	■
Metal 1	Metal 1	▨
Metal 2	Metal 2	▨

width is 0.1 μm and the value of X is 0.05 μm. The minimum line widths and spacings are determined primarily by the process technology and equipment used and especially the lithographic process. However, they are also determined in part by lateral doping and depletion effects. Implanted regions spread laterally during the annealing process, resulting in lateral

TABLE I.3

Simplified Layout Legend

Physical layer	Name	Layout symbol
n-well	n-well	□
Silicon nitride	Active	□
Polysilicon	Poly 1	░
p+ implant	p select	□
n+ implant	n select	□
Contact cut	Contact	■
Metal 1	Metal 1	□
Metal 2	Metal 2	□

doping. The diffusion of impurities also results in lateral doping effects. In addition, there are depletion regions surrounding implantations or diffusions made in a semiconductor of opposite conductivity type. Both the lateral doping and the depletion regions affect the minimum spacings of doped regions.

Violation of the minimum line width or spacing rules may result in a nonfunctioning circuit because of broken lines (if the minimum line width is violated) or a short circuit (if the minimum spacing between lines is violated).

Figures I.1 through I.5 illustrate the minimum widths and separations for polysilicon, implanted regions, and metal. The minimum width for polysilicon is 2X (rule *L1*) and the minimum polysilicon-polysilicon separation is also 2X (rule *D1*). The minimum width for implanted regions (3X, rule *L5*) is greater than for polysilicon to allow for depletion effects at the edges of the

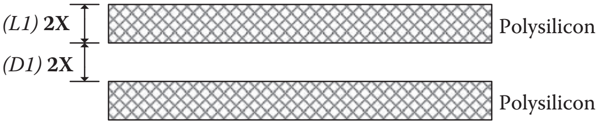


FIGURE I.1
Polysilicon design rules.

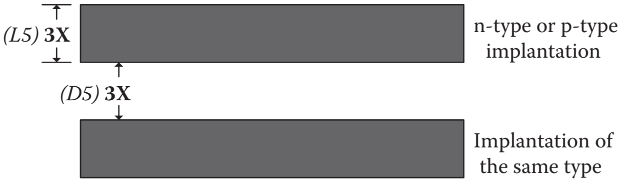


FIGURE I.2
Design rules for implantations of the same type.

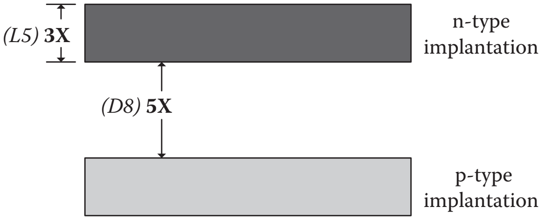


FIGURE I.3
Design rules for implantations of opposite type.

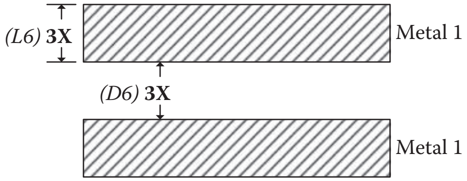


FIGURE I.4
Metal 1 design rules.

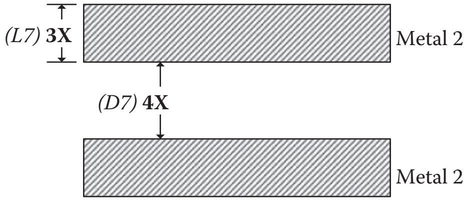
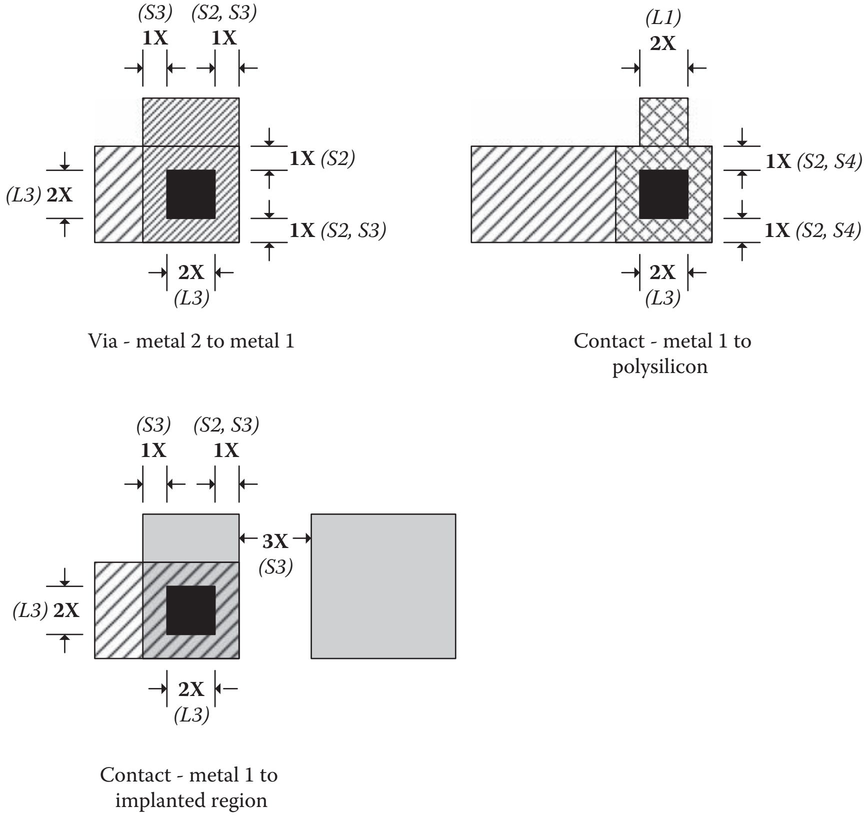


FIGURE I.5
Metal 2 design rules.

doped region. The minimum spacing design rule for implanted regions of opposite conductivity has been made large (5X, rule *D8*) to reduce the current gain of parasitic bipolar transistors and thereby avoid the latch-up problem. The minimum width for metal 1 is 3X (rule *L6*) and the minimum spacing for metal 1 is also 3X (rule *D6*). The widths and separations for higher levels of metal are generally greater, to allow for the loss of planarity on the surface as well as registration errors between mask levels.

Contacts are made to n+, p+, or polysilicon device regions by opening windows in the overlying oxide before metallization (see Figure I.6). In the scalable rule set adopted here, the minimum dimension for a metal contact is 2X (rule *L3*). 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 contact windows must be spaced by 2X (rule *D3*). The minimum surround for a contact window is 1X for metal 1 (rule *S2*), metal 2 (rule *S3*), polysilicon (rule *S4*), or an nselect or pselect region (rule *S5*). Therefore, the contacting layer as well as the layer being contacted must extend 1X (one-half the minimum feature size) in all directions, allowing for the tolerance in registration between the two mask levels. A special type of contact made between levels of metal is usually referred to as a *via*. An example is the case of a contact made between metal 1 and metal 2.

The basic design rules for a MOSFET are illustrated in Figure I.7 for the case of an n-MOS transistor. The channel is formed in the area in which the polysilicon wire overlaps the active region. The minimum width for this



Via - metal 2 to metal 1

Contact - metal 1 to polysilicon

Contact - metal 1 to implanted region

FIGURE I.6
Design rules for contacts and vias.

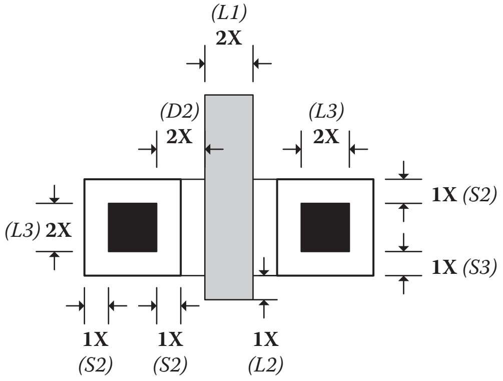


FIGURE I.7
n-MOS transistor design rules.

polysilicon wire, and therefore the "printed gate length L," is 2X (rule *L1*), and the polysilicon wire must extend beyond the active region by at least 1X on either side (rule *L2*). The nselect implant (not shown in these simplified layout drawings) must extend 1X (rule *S6*) beyond the edges of the active region. The contact windows are always 2X square (rule *L3*), but multiple contact openings are used if additional contact area is needed. The metal placed over a contact opening must extend beyond the window edges by 1X in all directions (rule *S2*), and this metal must be spaced by at least 1X from the channel region. 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.

In the case of a p-MOS transistor, an n-well surrounds the device as shown in Figure I.8. The n-well must extend at least 5X out from the active region in all directions (rule *S7*).

These scalable design rules apply for an n-well process, in which the n-MOS transistor is fabricated directly in the p-type substrate. In the case of a twin well process, a p-type well is created for the n-MOS device, so there are additional rules associated with the p-well dimensions. If absolute (rather than scalable) rules are used, they will apply to the same dimensions, separations, and extensions described above, but they will be in units of length rather than multiples of X.

Often it is necessary to connect two or more MOSFETs in series as shown in Figure I.9 for the case of two n-MOS transistors. In such a situation, it is not necessary to form source and drain contact regions between the series connected transistors; instead, the common nselect region between the two channels forms the drain of one transistor and the source of the other. The minimum separation between the two polysilicon gates is 2X (rule *D1*).

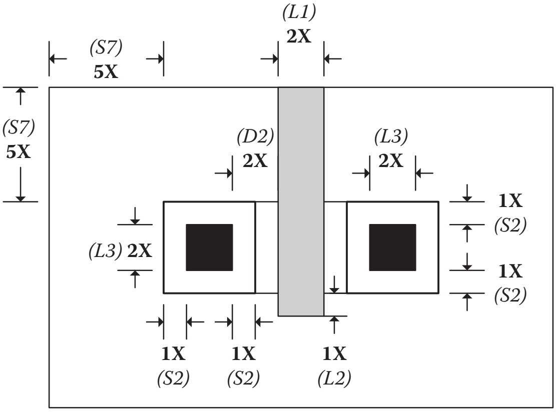


FIGURE I.8
p-MOS transistor design rules.

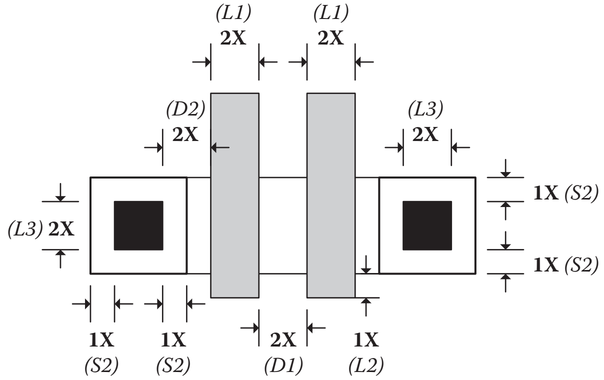


FIGURE I.9
Design rules for series-connected n-MOS transistors.

In many situations, it is also necessary to place two MOS transistors in parallel, to create the desired logic function or perhaps just to increase the overall current drive capability, as shown in Figure I.10. These two transistors share a single source region. Multiple contacts have been used for both drains as well as the common source. These contact areas are 2X in width (rule *L3*) and are spaced by 2X (rule *D3*).

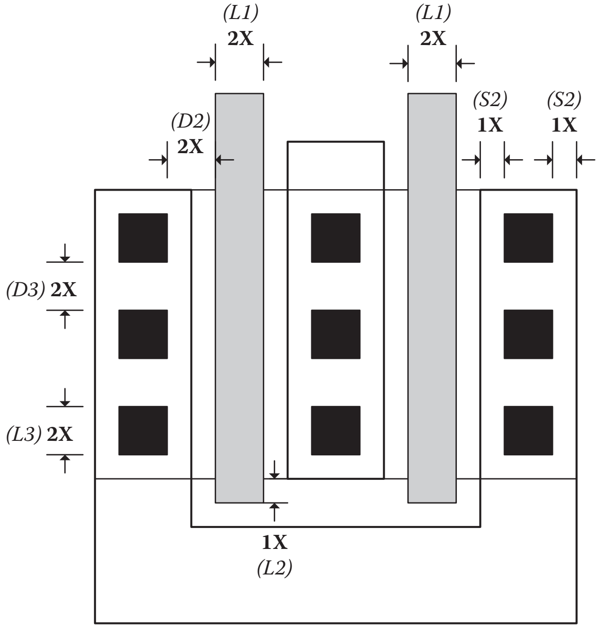


FIGURE I.10
Design rules for parallel-connected n-MOS transistors.

References

1. The MOSIS Service, Marina del Rey, CA, http://www.mosis.com.

Appendix J

p-n Junction Switching Transients

J.1 Introduction

For MOS VLSI circuits, the p-n junctions are normally reverse biased. Nonetheless, the large signal switching characteristics are important in several situations, including electrostatic discharge and latch-up. In this section, we will briefly consider the turn-on and turn-off transients for p-n junctions, revealing that the turn-on response is very rapid but that the turn-off response may be sluggish because of minority carrier charge storage effects.

J.2 Charge Control Model

The charge control equation for a one-sided n$^+$-p junction can be developed by consideration of the continuity equation for minority carriers in the p-type base. If we neglect generation and drift, this continuity equation is

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

Multiplying both sides of the equation by –q and integrating over the width of the base region, we obtain

$$-\frac{\partial}{\partial t}\int_0^{W_B} q n_p' dx = \int_0^{W_B} \frac{q n_p'}{\tau_n} dx + \int_0^{W_B} \frac{1}{A}\frac{\partial i_n(t)}{\partial x} dx . \tag{J.2}$$

Multiplying by the junction area yields

$$\frac{dQ_B(t)}{dt} + \frac{Q_B(t)}{\tau_n} - i_n\left(W_B, t\right) = -i_n\left(0, t\right) , \tag{J.3}$$

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 attributable to transit and recombination, so that

$$-i_n(0,t) = \frac{Q_B}{\tau_n} + \frac{Q_B}{t_{tB}} = \frac{Q_B}{\tau_F}, \tag{J.4}$$

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

$$-i_n(0,t) = \frac{dQ_B}{dt} + \frac{Q_B}{\tau_F}. \tag{J.5}$$

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

$$-i_n(0,t) = \frac{dQ_B}{dt} + \frac{Q_B}{\tau_F} - C_T\frac{dv}{dt}. \tag{J.6}$$

J.3 Turn-Off Transient

p-n junction junctions require a finite time to turn off attributable to two effects: the storage of excess minority carrier charge and the transition layer capacitance. Suppose a reverse voltage is applied after a junction has been forward biased for a long time as shown in Figure J.1. Before the bias is switched,

$$i(t) = I_F \approx \frac{V_F}{R}; \quad (t < 0). \tag{J.7}$$

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

$$i(t) = -I_R \approx -\frac{V_R}{R}; \quad (0 \le t \le t_s). \tag{J.8}$$

Application of the charge control equation yields

$$I_R \approx \frac{dQ_B}{dt} + \frac{Q_B}{\tau_F}. \tag{J.9}$$

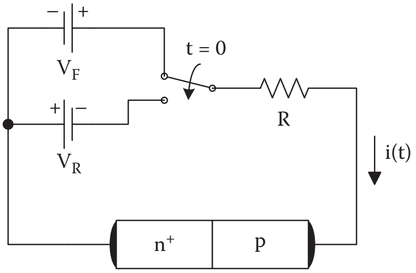


FIGURE J.1
p-n junction turn-off transient.

The solution is

$$Q_B(t) \approx I_R\tau_F - (I_R + I_F)\tau_F \exp(-t/\tau_F). \tag{J.10}$$

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 both transit and recombination effects. Solving, the storage delay time is

$$t_s = \tau_F\left[\ln\left(1 + \frac{I_F}{I_R}\right) - \ln\left(1 + \frac{\tau_R}{\tau_F}\right)\right]. \tag{J.11}$$

From this analysis, we can conclude that the actual switching speed of the junction depends on not only the device design (through τ_F and τ_R) but also on the circuit design (through I_F and I_R).

J.4 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 is suddenly forward biased as shown in Figure J.2. If the junction has been in a state of zero bias for a long time, the junction voltage cannot change abruptly attributable to the junction capacitance. Hence, 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

$$-I_F \approx \frac{dQ_B}{dt} + \frac{Q_B}{\tau_F}. \tag{J.12}$$

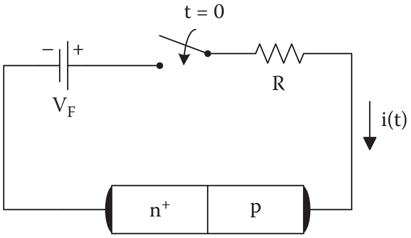


FIGURE J.2
p-n junction turn-on transient.

The initial and final conditions are

$$Q_B\left(t=0\right)=0\,, \tag{J.13}$$

and

$$Q_B\left(t=\infty\right)=-I_F\tau_F\,, \tag{J.14}$$

The solution is

$$Q_B(t)=-I_F\tau_F\left[1-\exp\left(-t/\tau_F\right)\right]. \tag{J.15}$$

We can determine the junction voltage as a function of time if we assume 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,

$$Q_B(t)\propto\bar{n}_p\exp\left(\frac{qv(t)}{kT}\right), \tag{J.16}$$

or

$$v(t)=V_{ss}-\frac{kT}{q}\ln\left(\frac{1}{1+\exp\left(-t/\tau_F\right)}\right), \tag{J.17}$$

where V_{ss} is the steady-state voltage. This result shows that both the current and voltage increase very rapidly when the junction is switched on.

Appendix K

Bipolar and BiCMOS Circuits

K.1 Introduction

Bipolar junction transistors have been used extensively in digital circuits because they exhibit higher current drive capability than MOSFETs having the same device area, making them well suited for driving large load capacitances. However, bipolar junction transistors are current-controlled devices, so bipolar circuits exhibit significant static dissipation, and this limits their packing density. Although bipolar transistor circuits remain important for linear and mixed-signal integrated circuits, CMOS is the dominant technology for digital circuits and is therefore emphasized in this book. The purpose of this appendix is to provide a concise reference on bipolar junction transistors and some digital circuits that use them, including TTL, Schottky TTL, ECL, and BiCMOS circuits.

K.2 Bipolar Junction Transistors

A BJT is a three-terminal device comprising two p-n junctions. Both n-p-n and p-n-p bipolar transistors may be fabricated, but n-p-n transistors are preferred because of their superior characteristics. Figure K.1 illustrates an n-p-n bipolar transistor with its circuit symbol. The p-region is the base (B) and forms p-n junctions with the emitter (E) and collector (C). In the forward active mode of operation, the forward-biased emitter-base junction injects minority carriers (electrons) that diffuse across the base and are collected by the reverse-biased collector-base junction. Therefore, in this mode of operation, the bipolar transistor acts as a current-controlled current source, with the base current I_B controlling the collector current I_C.

There are four modes of operation for the BJT: *cutoff*, *forward active*, *reverse active*, and *saturation*. In a typical digital circuit, the bipolar transistor is

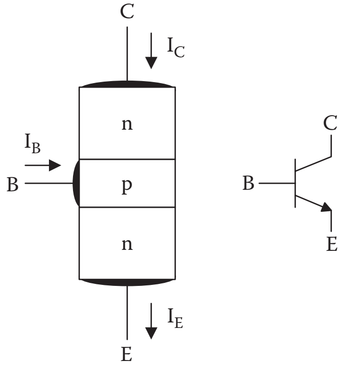


FIGURE K.1
n-p-n bipolar transistor and circuit symbol.

used as a switch that operates in the cutoff (“off”) or saturation (“on”) mode. However, all four modes of operation are used in digital bipolar circuits, and a switch transistor passes through forward active operation when switching between the cutoff and saturation modes.

Figure K.2 shows typical characteristics for an n-p-n bipolar junction transistor. Figure K.2a shows the collector current versus 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 K.2b, 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

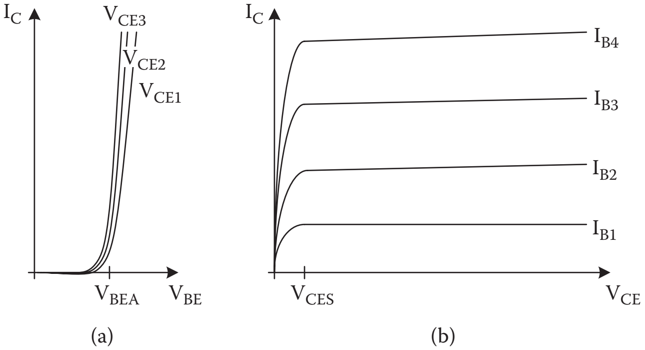


FIGURE K.2
Typical characteristics for an n-p-n bipolar junction transistor.

$V_{CE} \leq V_{CES}$. 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 $V_{CE} < 0$.

K.2.1 Cutoff Operation

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

$$I_C \approx I_E \approx I_B \approx 0 \ \text{(cutoff operation).} \tag{K.1}$$

K.2.2 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 $V_{BC} < 0$). 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. There is a linear profile for the excess minority carrier concentration in the base as shown in Figure K.3. This is analogous to the situation in a short-base diode, and such a transistor is called a short-base transistor.

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. It is

$$n_B(0) - \bar{n}_B \approx \bar{n}_B\left(e^{qV_{BE}/kT} - 1\right), \tag{K.2}$$

where n_B is electron concentration in the base, and $\bar{n}_B$ is equilibrium electron concentration in the base. This results in an approximately linear profile for the excess minority carrier concentration in the base (see Figure K.3) given by

$$n_B - \bar{n}_B \approx \bar{n}_B\left(e^{qV_{BE}/kT} - 1\right)\left(\frac{W_B - x}{W_B}\right), \tag{K.3}$$

where x is the distance from the emitter, and W_B is the base width. The resulting current attributable to the diffusing minority carriers is

$$I_C = qAD_{nB}\frac{dn_B}{dx} \approx \frac{qAD_{nB}n_i^2}{W_B N_{aB}}\left(e^{qV_{BE}/kT} - 1\right), \tag{K.4}$$

where q is electronic charge, A is emitter junction area, D_{nB} is diffusivity of electrons in the base, n_i is intrinsic carrier concentration in Si, W_B is base width, and N_{aB} is acceptor concentration in the base. Therefore,

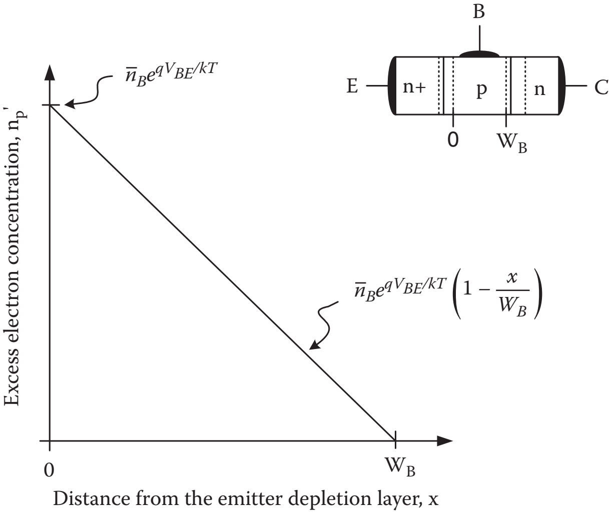


FIGURE K.3
Excess minority carrier profile in the base of an n-p-n bipolar transistor operating in the forward active mode.

the collector current increases exponentially with the base-emitter bias voltage.

The common base current gain α_F of the bipolar transistor is

$$\alpha_F \approx \gamma_E \alpha_T , \tag{K.5}$$

where the emitter injection efficiency γ_E is

$$\gamma_E = \left(1 + \frac{D_{pE} N_{aB} W_B}{D_{nB} N_{dE} L_{pE}}\right)^{-1} , \tag{K.6}$$

and the base transport factor α_T is given by

$$\alpha_T = \left\{\cosh\left(W_B / L_{nB}\right)\right\}^{-1} \approx \left(1 + \frac{W_B^2}{2L_{nB}^2}\right)^{-1} . \tag{K.7}$$

The common emitter current gain β_F is related to the common base current gain by

$$\beta_F = \frac{\alpha_F}{1 - \alpha_F} . \tag{K.8}$$

For hand calculations, forward active operation is modeled using

$$V_{BE} \approx V_{BEA} \text{ and } I_C \approx \beta_F I_B \text{. (forward active operation)} \tag{K.9}$$

K.2.3 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 $V_{BC} > 0$). Here,

$$V_{BC} \approx V_{BCA} \text{ and } I_E \approx \beta_R I_B \text{, (reverse active operation)} \tag{K.10}$$

where β_R is the reverse common-emitter current gain.

K.2.4 Saturation Operation

In the saturation mode of operation, both junctions are forward biased ($V_{BE} > 0$ and $V_{BC} > 0$). Both forward biased junctions 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

$$V_{BE} \approx V_{BES} \text{ and } V_{CE} \approx V_{CES} \text{ (saturation operation).} \tag{K.11}$$

K.2.5 Bipolar Transistor SPICE Model

In SPICE, the device equations are based on the Gummel-Poon model and the circuit diagram shown in Figure K.4.

I_{BE} and I_{CB} are Schockley-type current sources including adjustable emission coefficients. These currents are calculated by

$$I_{CB} = IS\left[\exp\left(\frac{qV_{BE}}{NFkT}\right) - \exp\left(\frac{qV_{BC}}{NRkT}\right)\right]\left[1 - \frac{V_{BC}}{VAF}\right] - \frac{IS}{BR}\left[\exp\left(\frac{qV_{BC}}{NRkT}\right) - 1\right], \tag{K.12}$$

and

$$I_{BE} = IS\left[\exp\left(\frac{qV_{BE}}{NFkT}\right) - \exp\left(\frac{qV_{BC}}{NRkT}\right)\right]\left[1 - \frac{V_{BC}}{VAF}\right] + \frac{IS}{BF}\left[\exp\left(\frac{qV_{BE}}{NFkT}\right) - 1\right], \tag{K.13}$$

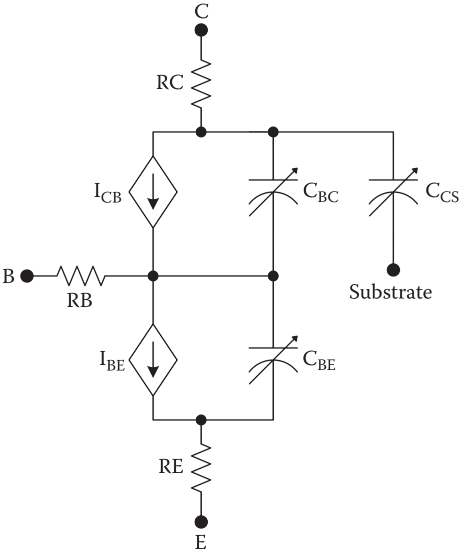


FIGURE K.4
SPICE circuit model for the bipolar junction transistor.

where V_{BE} is base-emitter voltage, V_{BC} is base-collector voltage, *IS* is junction saturation current, *NF* is forward emission coefficient, *NR* is reverse emission coefficient, *BF* is forward beta, *BR* is reverse beta, *VAF* is forward Early voltage, and ϕ_T is thermal voltage (26 mV at 300 K). C_{BE} and C_{BC} are the base-emitter and base-collector capacitances, respectively, and include both the depletion and diffusion contributions. C_{CS} is the collector-substrate capacitance, which is a depletion layer capacitance. These capacitances are calculated by

$$C_{BE} = TF\frac{IS}{NF\phi_T}\exp\left(\frac{qV_{BE}}{NFkT}\right) + \frac{CJE}{\left(1-\frac{V_{BE}}{VJE}\right)^{MJE}}, \tag{K.14}$$

$$C_{BC} = TR\frac{IS}{NR\phi_T}\exp\left(\frac{qV_{BC}}{NRkT}\right) + \frac{CJC}{\left(1-\frac{V_{BC}}{VJC}\right)^{MJC}}, \text{ and} \tag{K.15}$$

$$C_{CS} = \frac{CJS}{\left(1-\frac{V_{CS}}{VJS}\right)^{MJS}}, \tag{K.16}$$

where V_{BE} is base-emitter voltage, V_{BC} is base-collector voltage, *IS* is junction saturation current, *NF* is forward emission coefficient, *NR* is reverse emission coefficient, CJE is zero-bias base-emitter capacitance, CJC is zero-bias base-collector capacitance, VJE is base-emitter built-in potential, VJC is base-collector built-in potential, VJS is collector-substrate built-in potential, MJE is base-emitter grading coefficient, MJC is base-collector grading coefficient, and MJS is collector-substrate grading coefficient. RC, RB, and RE are the series resistances in the device.

K.3 Bipolar Transistor Inverter and the Saturation Delay

A dominant time delay component in a TTL-type bipolar logic circuit is the output transistor saturation delay. This delay may be most easily understood by consideration of the simple bipolar transistor inverter from which TTL evolved.

First, consider a bipolar transistor inverter for which the input makes an abrupt low-to-high transition as shown in Figure K.5. There are two important contributions to the high-to-low propagation delay. These are the delay time for cutoff operation and the fall time. The delay time is

$$t_D = \frac{V_{BEA}\left(C_{BE} + C_{BC}\right)}{I_B(ave)}, \tag{K.17}$$

where C_{BE} and C_{BC} are the average junction capacitances, and I_B(ave) is the average base current. The junction depletion capacitances C_{BE} and C_{BC} depend on the base-emitter and base-collector bias voltages, respectively, and are both time dependent. For the purpose of hand calculations, average junction depletion capacitance values are used. Thus, if the junction voltage varies from an initial value V_1 to a final value of V_2, the average junction capacitance is

$$C_J = \frac{C_{JO}V_{bi}}{\left(V_1 - V_2\right)\left(1 - m\right)}\left[\left(1 - \frac{V_2}{V_{bi}}\right)^{1-m} - \left(1 - \frac{V_1}{V_{bi}}\right)^{1-m}\right], \tag{K.18}$$

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 output fall time is approximately

$$t_F = \frac{I_{CEOS}\tau_F + \Delta V_{BC}C_{BC}}{I_B(ave)}. \tag{K.19}$$

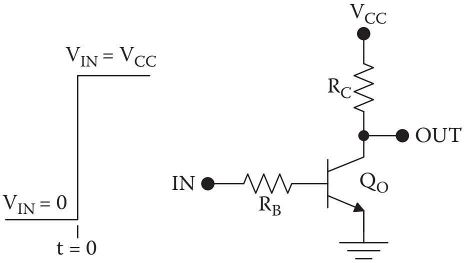


FIGURE K.5
Bipolar transistor inverter for consideration of t_{PHL}.

where I_{CEOS} is the collector current at the edge of saturation, τ_F is the forward transit time for minority carriers in the base, ΔV_{BC} is the change in the base-collector voltage during the fall time, C_{BC} is the average base-collector depletion capacitance, and $I_B(ave)$ is the average base current. The high-to-low propagation delay time is

$$t_{PHL} = t_D + \frac{t_F}{2}. \tag{K.20}$$

Now suppose the input voltage has been high for a long time and makes an abrupt high-to-low transition as shown in Figure K.6. There are two contributions to the low-to-high propagation delay: the saturation delay and the rise time. The saturation delay is

$$t_S = \tau_S \ln\left(\frac{I_{BF} - I_{BR}}{I_{CEOS} / \beta_F - I_{BR}}\right), \tag{K.21}$$

where τ_S is the saturation time constant, I_{BF} is the forward base current during saturation operation, and I_{BR} is the base current while the transistor is being brought out of saturation. The saturation time constant is

$$\tau_S = \frac{\alpha_F\left(\tau_F + \alpha_R \tau_R\right)}{1 - \alpha_F \alpha_R}, \tag{K.22}$$

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. The rise time is given by

$$t_R = \frac{I_{CEOS}\tau_F + \left|\Delta V_{BC} C_{BC}\right|}{\left|I_B(ave)\right|}, \tag{K.23}$$

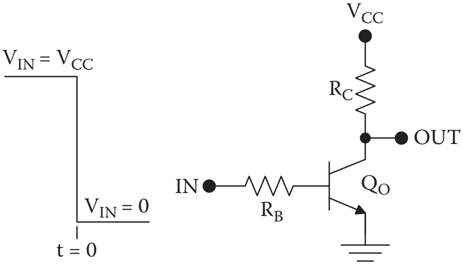


FIGURE K.6
Bipolar transistor inverter for consideration of t_{PLH}.

and the low-to-high propagation delay is

$$t_{PLH} = t_S + \frac{t_R}{2}\,. \tag{K.24}$$

Typically, the saturation delay on the order of 10 ns, much longer than the other delay terms. Therefore, high-speed bipolar and BiCMOS circuits are designed to avoid saturated operation of the bipolar transistors.

K.4 Transistor-Transistor Logic Circuits

TTL circuits are based on the simple bipolar transistor inverter but have additional components to provide improved voltage swing, fan-out, and transient response. TTL circuits use active pull-up and active pull-down for fast response with a large load capacitance. However, several of the transistors in a TTL circuit are allowed to saturate. The saturation delay for the output pull-down transistor is an important speed limitation for standard TTL circuits. Schottky TTL circuits have been developed to avoid this problem. These circuit families use Schottky clamp diodes on the bipolar transistors to prevent saturation and therefore obtain shorter propagation delays than standard TTL. Nonetheless, TTL circuits exhibit relatively high static dissipation so their power-delay products are inferior to those for modern CMOS gates.

A standard TTL inverter circuit is shown in Figure K.7. With a logic "0" input, Q_I is saturated, resulting in a small voltage at the base of Q_S. Therefore, both Q_S and Q_O are cutoff, Q_P is forward active, and the output goes high. The output high voltage is less than V_{CC} because of the base-emitter drop in Q_P and the diode voltage drop in D_O; hence

$$V_{OH} \approx V_{CC} - V_{BEA} - V_D \approx 3.4V. \tag{K.25}$$

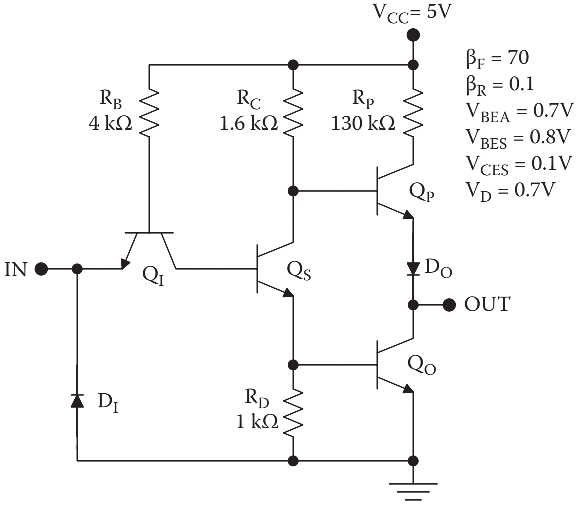


FIGURE K.7
Standard TTL inverter.

With a logic "1" input, Q_I is reverse active and provides base drive to Q_S. Both Q_S and Q_O saturate, so that

$$V_{OL} \approx 0.1V. \tag{K.26}$$

A standard TTL NAND gate may be constructed by implementing multiple emitters in the input transistor, as shown in Figure K.8. In this circuit, the

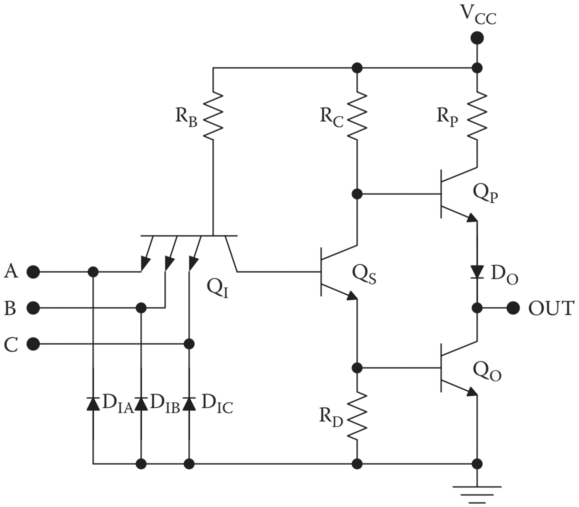


FIGURE K.8
Standard TTL NAND3 gate.

input transistor saturates if one or more of the inputs goes low, providing the NAND function. For the NOR function, multiple input transistors are used, as shown in Figure K.9 for the case of two inputs.

Because several of the transistors in the standard TTL circuit are allowed to saturate, saturation delays represent an important speed limitation, and the low-to high propagation delay t_{PLH} is dominated by the saturation delay for Q_O. This led to the development of TTL circuit families with Schottky-clamped bipolar transistors as shown in Figure K.10.

In the Schottky-clamped bipolar transistor, a Schottky diode is placed between the base and collector. In silicon technology, a typical Schottky diode has a turn-on voltage of 0.3 V. Therefore, the Schottky diode will shunt excess base drive current to the collector, preventing the base-collector junction from reaching a forward bias voltage of $V_{BCA} \approx 0.7V$ and preventing saturation.

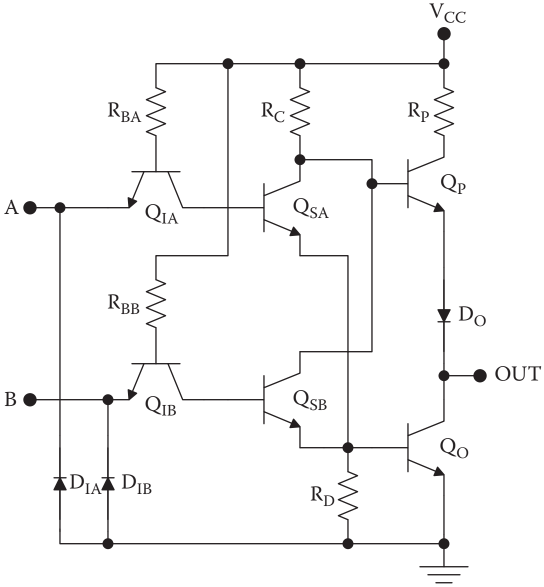


FIGURE K.9
Standard TTL NOR2 circuit.

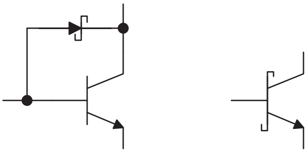


FIGURE K.10
Schottky-clamped transistor (left) and circuit symbol (right).

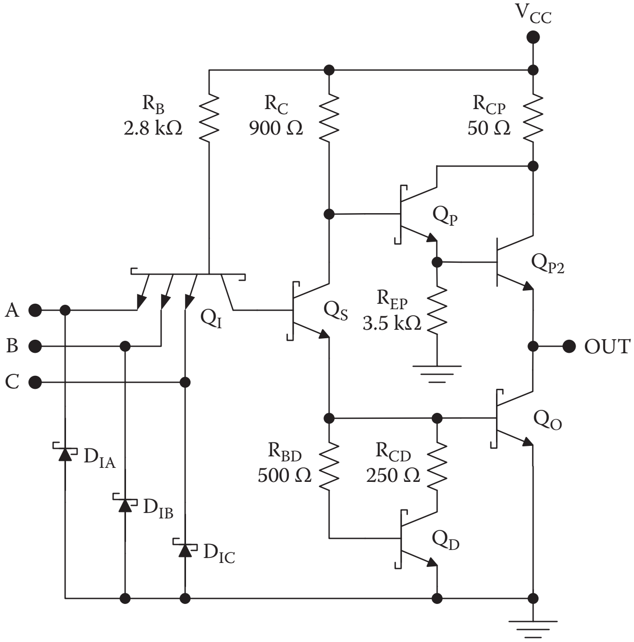


FIGURE K.11
Schottky TTL NAND3 circuit.

Figure K.11 depicts a Schottky TTL (STTL family) inverter, and Figure K.12 shows a low-power Schottky (LSTTL family) NAND gate. In the LSTTL circuits, the multi-emitter input transistor has been replaced by Schottky diodes.

K.5 Emitter-Coupled Logic Circuits

ECL circuits achieve superior delay characteristics compared with other bipolar logic circuits through the use of a small voltage swing, the avoidance of saturation in the bipolar transistors, and a low-impedance emitter follower to drive the output. On the other hand, ECL circuits suffer from several disadvantages compared with CMOS, including high-power dissipation, use of a negative voltage supply, and complex circuits with low packing density.

A basic ECL inverter/buffer circuit is illustrated in Figure K.13. This circuit uses a negative supply voltage of –5.2 V. Two bipolar transistors Q_I and Q_R are arranged in a differential pair and switch a nearly constant current I_X between their collector resistors R_{CI} and R_{CR}. Q_{INV} and Q_{NINV} are low-impedance emitter followers for the complementary outputs. Two ground

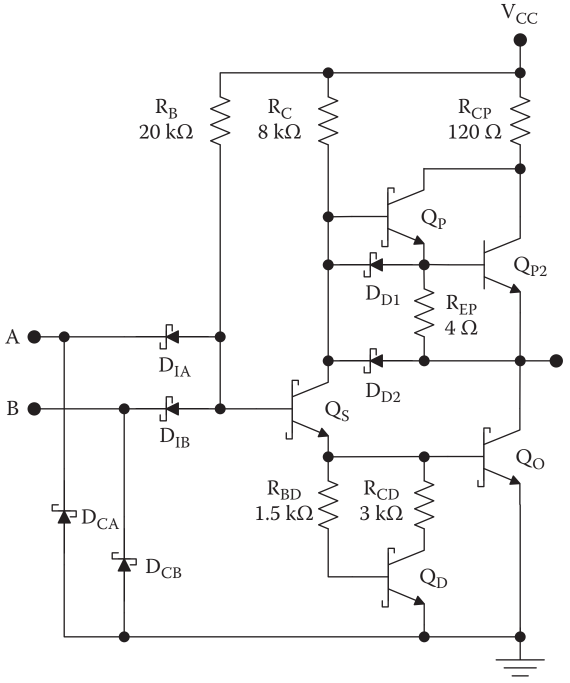


FIGURE K.12
Low-power Schottky TTL NAND2 circuit.

connections are used. V_{CC1} is the "dirty ground," which exhibits sizable voltage glitches during switching events. The output levels are referenced to V_{CC2}, the "clean ground," which is electrically quiet.

The behavior of the inverting output is as follows. When $V_{IN} < V_{REF}$, Q_I is cutoff and Q_R carries essentially all of the current I_X. There is negligible current (and voltage drop) in R_{CI}, but the emitter follower is forward active so there is a voltage drop of V_{BEA} between its base and the inverting output. Thus,

$$V_{OH} \approx -V_{BEA}. \tag{K.27}$$

When $V_{IN} > V_{REF}$, Q_I is forward active, whereas Q_R is cutoff. There is a voltage drop equal to $I_X R_{CI}$ in the resistor R_{CI}, so that

$$V_{OL} \approx -I_X R_{CI} - V_{BEA} \approx -2V_{BEA}. \tag{K.28}$$

Therefore, the voltage swing is approximately V_{BEA}. (For a silicon ECL circuit, V_{BEA} is usually assumed to be 0.75 V attributable to the high current densities

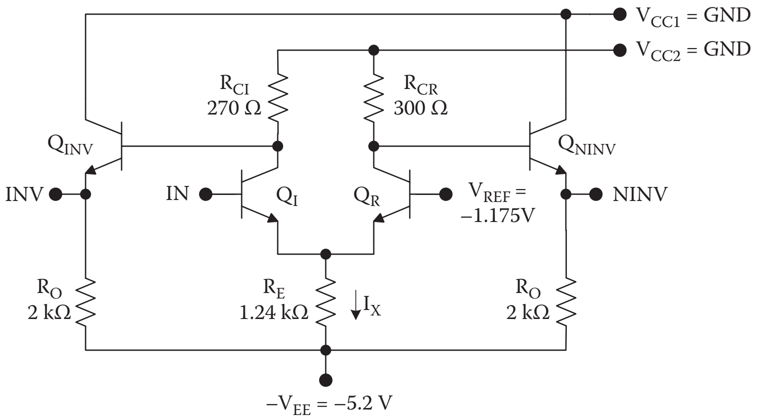


FIGURE K.13
ECL inverter/buffer circuit.

used in the bipolar transistors.) The behavior of the complementary output is qualitatively similar, with approximately the same values of V_{OH} and V_{OL}.

The critical voltages V_{IL} and V_{IH} may be estimated by assuming a 100 mV transition region at the input. Thus,

$$V_{IL} \approx V_{REF} - 0.05V \tag{K.29}$$

and

$$V_{IH} \approx V_{REF} + 0.05V\,. \tag{K.30}$$

An ECL circuit exhibits high static power dissipation attributable to the steady DC I_X. Neglecting the currents in the output resistors (R_O in Figure K.13), the static dissipation is approximately

$$P_{DC} \approx I_X V_{EE}\,, \tag{K.31}$$

which is usually 10–100 mW.

Analysis of the transient response for an ECL circuit is quite complex. However, the propagation delay may be estimated by considering a first-order RC circuit comprising R_{CI} and the effective capacitance appearing at the node connected to the base of the emitter follower Q_{INV}. By this analysis, the propagation delays may be estimated for a two-input ECL OR/NOR gate as

$$t_{PLH} \approx t_{PHL} \approx t_P \approx 5R_{CI}C_{BC}\,, \tag{K.32}$$

where C_{BC} is the base-collector capacitance for the bipolar transistors (assumed to be constant). Therefore, the delay times can be improved by

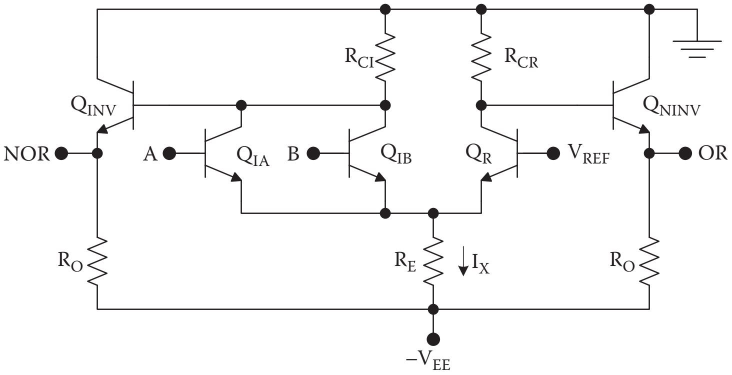


FIGURE K.14
Two-way ECL OR/NOR gate.

scaling down the resistors, but with a corresponding increase in the static dissipation.

Figure K.14 illustrates a two-way ECL OR/NOR gate. Here, if either of the inputs goes high, the associated input transistor is forward active and carries the current I_X, causing the inverting (NOR) output to go low.

There are a number of variations on the basic ECL circuits that were discussed above. Some use a current source for the emitter-coupled transistors for better temperature stability, active pull-down for the outputs for improved dynamic response, or a scaled supply voltage for reduced dissipation.

K.6 BiCMOS Logic Circuits

CMOS is preferred for most applications because of its low standby power, high packing density, and the speed performance improvements that can be achieved by device scaling. On the other hand, bipolar transistors exhibit greater transconductance than MOSFETs taking up the same silicon area. Therefore, bipolar transistor circuits can be configured for low output impedance to drive large load capacitances, such as those encountered with off-chip loads. BiCMOS circuits have been used to combine the low standby power of CMOS with the low output impedance of bipolar circuitry, primarily for off-chip or high fan-out loads exhibiting large capacitances (>10 pF).

Figure K.15 illustrates a BiCMOS inverter circuit, in which the MOS transistors M_{P1} and M_{N1} perform the logic function, bipolar transistors Q_P and Q_O are the output drivers, and the MOS transistors M_{N3} and M_{N2} help discharge base charge from the bipolar transistors for improved transient response.

The basic operation of the BiCMOS inverter is as follows. With a logic "1" input, M_{P1} is cutoff, but M_{N1} and M_{N3} are linear. The voltage at the

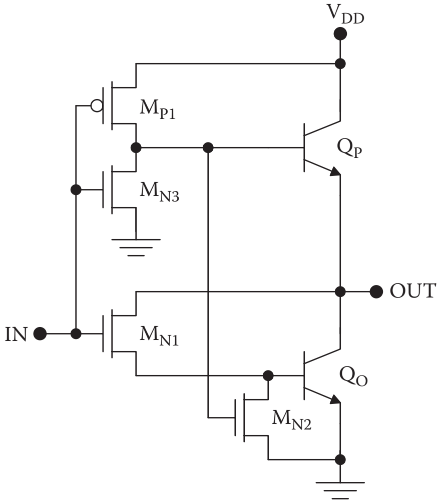


FIGURE K.15
BiCMOS inverter circuit.

base of Q_P goes to zero so Q_P is cutoff and M_{N2} is cutoff. Q_O will be forward active with a base-emitter voltage drop of V_{BEA}. Neglecting the voltage drop across the linear transistor M_{N1}, which carries negligible drain current,

$$V_{OL} \approx V_{BEA}\,. \tag{K.33}$$

With a logic "0" input, M_{P1} is linear, whereas M_{N3} and M_{N1} are cutoff. The voltage at the base of Q_P goes to V_{DD}. Therefore, M_{N2} is linear, shorting the base-emitter junction of Q_O and causing this transistor to be cutoff. The forward active bipolar transistor Q_P will therefore act as an emitter follower, with an output voltage equal to

$$V_{OH} \approx V_{DD} - V_{BEA}\,. \tag{K.34}$$

An important disadvantage of BiCMOS is its reduced voltage swing. Whereas CMOS exhibits rail-to-rail swing, the logic swing of BiCMOS is

$$LS = V_{OH} - V_{OL} \approx V_{DD} - 2V_{BEA}\,. \tag{K.35}$$

For typical silicon bipolar transistors, $V_{BEA} \approx 0.7V$ so the voltage swing is about 1.4 V less than the supply voltage. As a consequence, BiCMOS cannot be used in low-voltage, low-power applications.

To compare the transient response of CMOS and BiCMOS, consider the circuits shown in Figure K.16. Suppose that similar MOSFETs have been

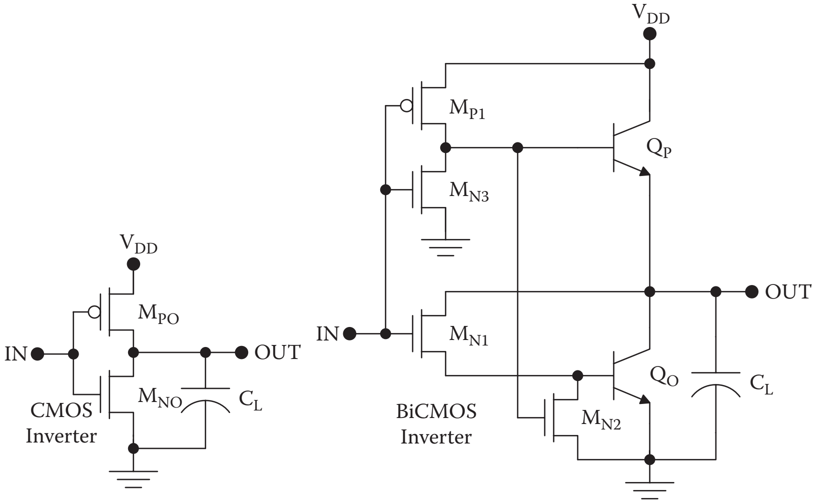


FIGURE K.16
CMOS and BiCMOS inverters with lumped capacitive loads.

used in both types of circuits with $K_N = K_P = K$ and $V_{TN} = |V_{TP}| = V_T$. For the symmetric CMOS circuit, $t_{PLH} = t_{PHL} = t_P$, and

$$t_P\left(CMOS\right) = \frac{C_L}{K}\Gamma \tag{K.36}$$

where

$$\Gamma = \frac{1}{\left(V_{DD} - V_T\right)}\left[\frac{2V_T}{\left(V_{DD} - V_T\right)} + \ln\left(\frac{3V_{DD} - 4V_T}{V_{DD}}\right)\right]. \tag{K.37}$$

The transient behavior of the BiCMOS circuit is quite complex, and the (unequal) propagation delays are best determined by SPICE. However, we can make a first-order estimate of both propagation delays as the sum of the two dominant time delay components. The first is the propagation delay for the M_{P1}-M_{N3} CMOS inverter loaded by the Q_P emitter follower. The load capacitance seen by this CMOS inverter is approximately equal to the base-collector capacitance of Q_P. The second time delay 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 the Q_P emitter follower ($1 / g_m$ is of the order of a few Ohms). Adding these two components , the propagation delays may be estimated as

$$t_P\left(BiCMOS\right) \approx \frac{C_{BC}}{K}\Gamma + \ln(2)\frac{C_L}{g_m}. \tag{K.38}$$

In many cases, it may be possible to neglect the second term. For example, with $1/g_m = 10\Omega$ and $C_L = 15pF$, the second term is 0.1 ns, whereas an off-chip propagation delay might be 100 times greater. If the second term is negligible, so that the internal propagation delay of the BiCMOS circuit dominates, then

$$\frac{t_P(CMOS)}{t_P(BiCMOS)} \approx \frac{C_L}{C_{BC}}. \tag{K.39}$$

It is therefore beneficial to use BiCMOS if $V_{DD}/2 > 2V_{BEA}$, so that a voltage swing of at least $V_{DD}/2$ is maintained, and $C_L > C_{BC}$, so that improved transient response can be achieved. These two conditions are met for high-voltage off-chip driver circuits.

Logic design in BiCMOS circuits involves replacing M_{N1} and M_{N3} with n-MOS logic networks while replacing M_{P1} by a dual p-MOS logic network. Figure K.17 shows a three-way BiCMOS NAND gate, and

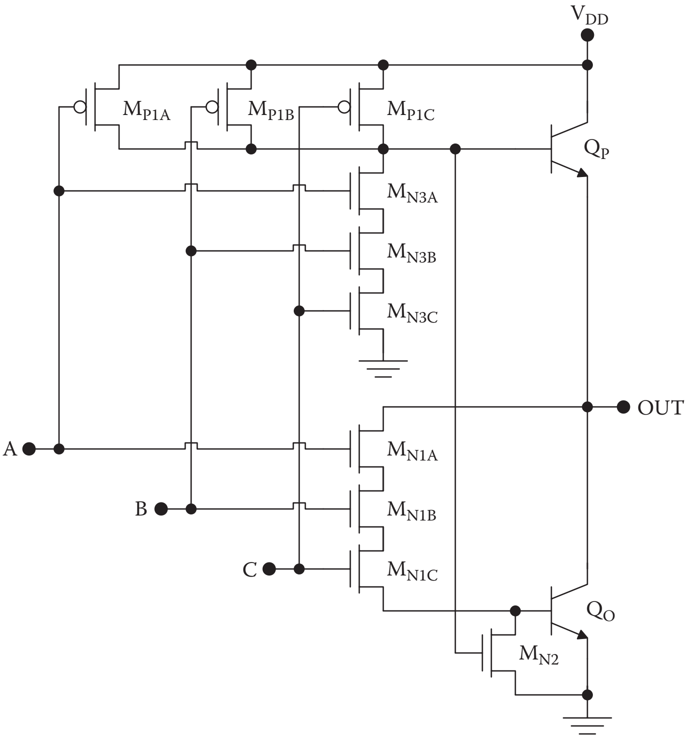


FIGURE K.17
BiCMOS NAND3 circuit.

Figure K.18 displays a two-way BiCMOS NOR gate. More complex logic functions can also be implemented, and the device scaling is governed by the same principles as in CMOS circuits. In general, a BiCMOS gate with fan-in M contains 3M + 1 MOS transistors plus two bipolar transistors.

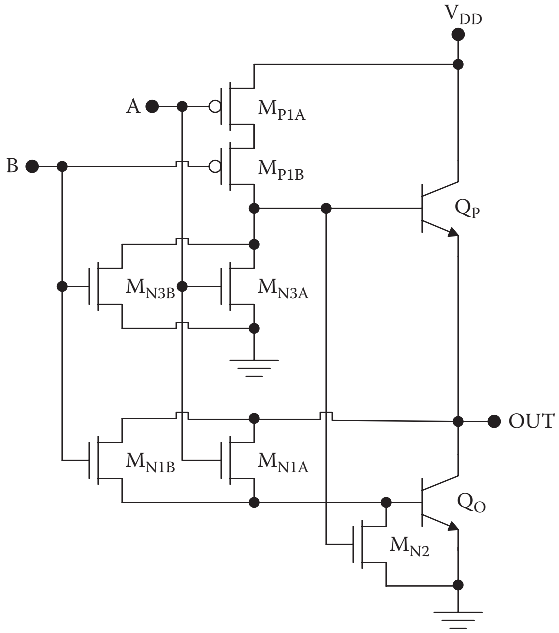


FIGURE K.18
BiCMOS NOR2 circuit.

Appendix L

Integrated Circuit Packages

L.1 Introduction

Once digital integrated circuits have been designed and fabricated, 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 of electrical connections, capability of high input and output data rates, and the 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 appendix is to provide a concise reference on the principles of integrated circuit packages.

L.2 Package Types

There are five basic types of integrated circuit packages [1]:

- **Through-hole packages (THTs)** have metal pins that may be inserted through holes drilled in the circuit board for soldering. This package technology has been around the longest but is inefficient in the use of printed circuit board area.
- **Surface mount technology packages** use metal leads that can be soldered to a single surface of the printed circuit board. They are much smaller and lighter weight than through-hole packages, for a given number of electrical connections. In addition, they are more resistant to mechanical shock compared with through-hole parts. Surface mount packages are growing in popularity for these reasons. In fact, some product applications (such as notebook computers, digital

* Although the plural of "die" is "dice," it is standard practice in industry to use "die" as the plural.

wireless devices, and personal media players) would not have been made possible without surface mount components.

- **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. On the other hand, chip-scale packages offer advantages in handling and testability compared with bare die. Usually, chip-scale packages are attached to circuit boards via an array of metal bumps, providing a high pin density and mechanical robustness.
- **Bare die, or *unpackaged* parts,** offer the minimum size and weight and also eliminate the 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 therefore introduce another level of packaging between the integrated circuit and the circuit board. Some modules use stacked die to achieve the minimum connection lengths and the highest efficiency in circuit board use.

Today there is an almost endless variety of integrated circuit packages. Some standards* have been established (for example, by the Joint Electron Device Engineering Council). However, manufacturers are introducing new packages at an ever-increasing rate. Some of these packages 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' websites for up-to-date information on package types.

L.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. Three important types are dual in-line packages (DIPs), quad in-line packages (QIPs), and pin grid arrays (PGAs). DIPs are rectangular packages with metal pins arranged along two sides; an example is illustrated in Figure L.1. QIPs have pins arranged along all four sides of the package for higher efficiency. PGAs use 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 L.2.

DIPs are by far the most popular THT packages and come in a number of varieties. Plastic DIPs are the most cost effective, whereas ceramic DIPs are

* Many integrated circuit package standards are based on the English system of units for historical reasons; therefore, pin spacings are sometimes specified in mils (1000 mil = 1 inch).

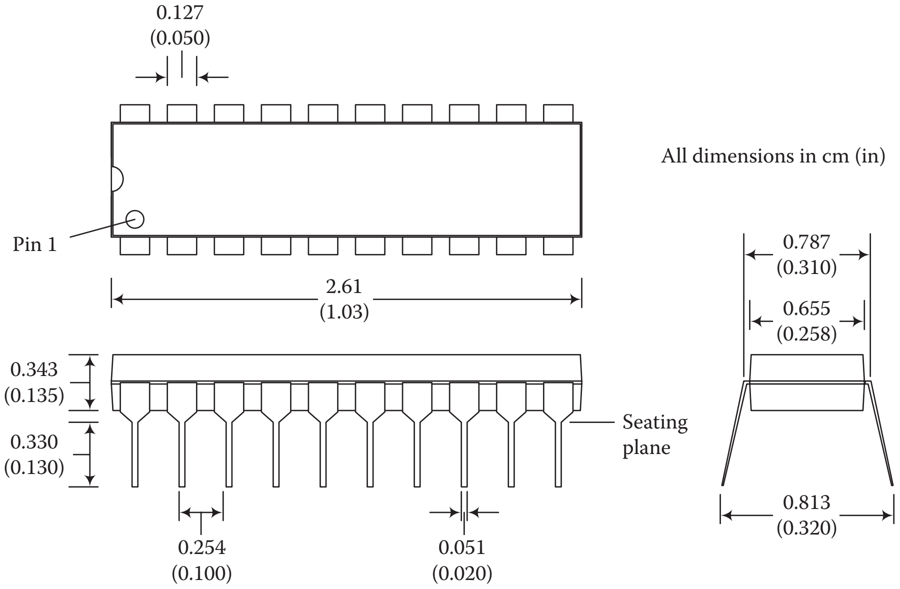


FIGURE L.1
Plastic dual in-line package with 20 pins.

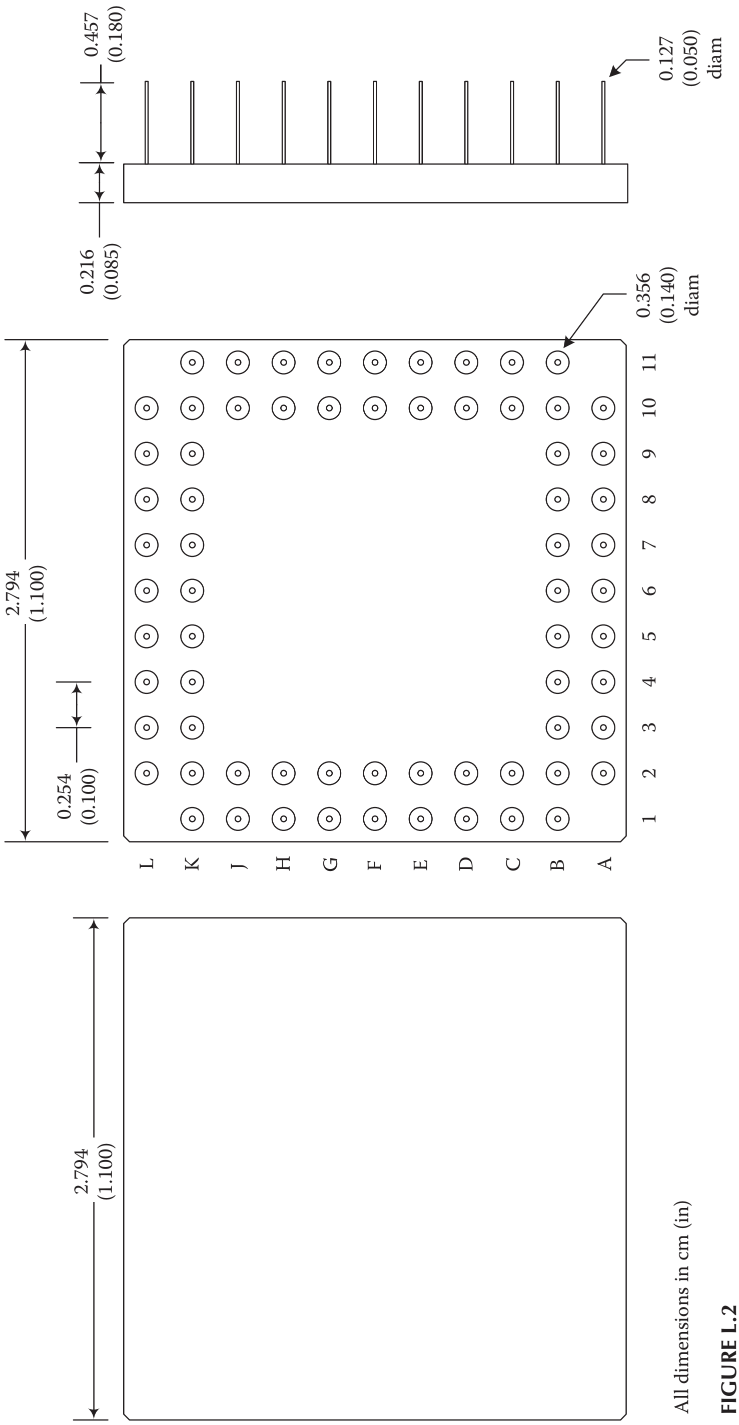
more suitable for high-power, high-temperature applications. Shrink DIPs (also known as skinny DIPs or SK-DIPs) use closer lead spacing and are more compact. Zig-zag in-line packages achieve even closer lead spacings in two zig-zag patterns.

QIPs (also known as QUIPs) use leads on all four sides. This advantage is slight compared with 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. Both plastic and ceramic versions are available.

L.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. However, the molding process may bring 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, J-leaded chip carriers, 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.



All dimensions in cm (in)

FIGURE L.2
Plastic PGA package with 68 pins.

J-lead chip carriers have leads that are J-shaped and 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 surface mount technology packages, the most popular are variations of the BGA, SOIC, and the leaded chip carrier (such as the plastic J-lead chip carrier). Several important types of surface mount pages are depicted in Figures L.3 through L.6.

L.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 with bare die. Chip scale packaging technologies include the popular micro ball grid array package styles. Nearly all chip scale packages use flip chip technology. Thus, the die is mounted top down on a ceramic substrate. Before die mounting, the aluminum pads on the die are built up with metal bumps. These attach one-to-one 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.

L.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 technology

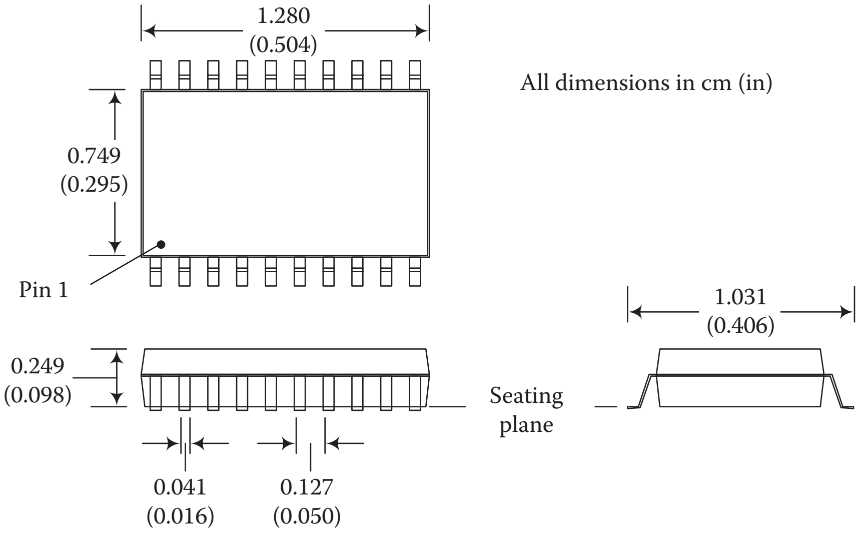
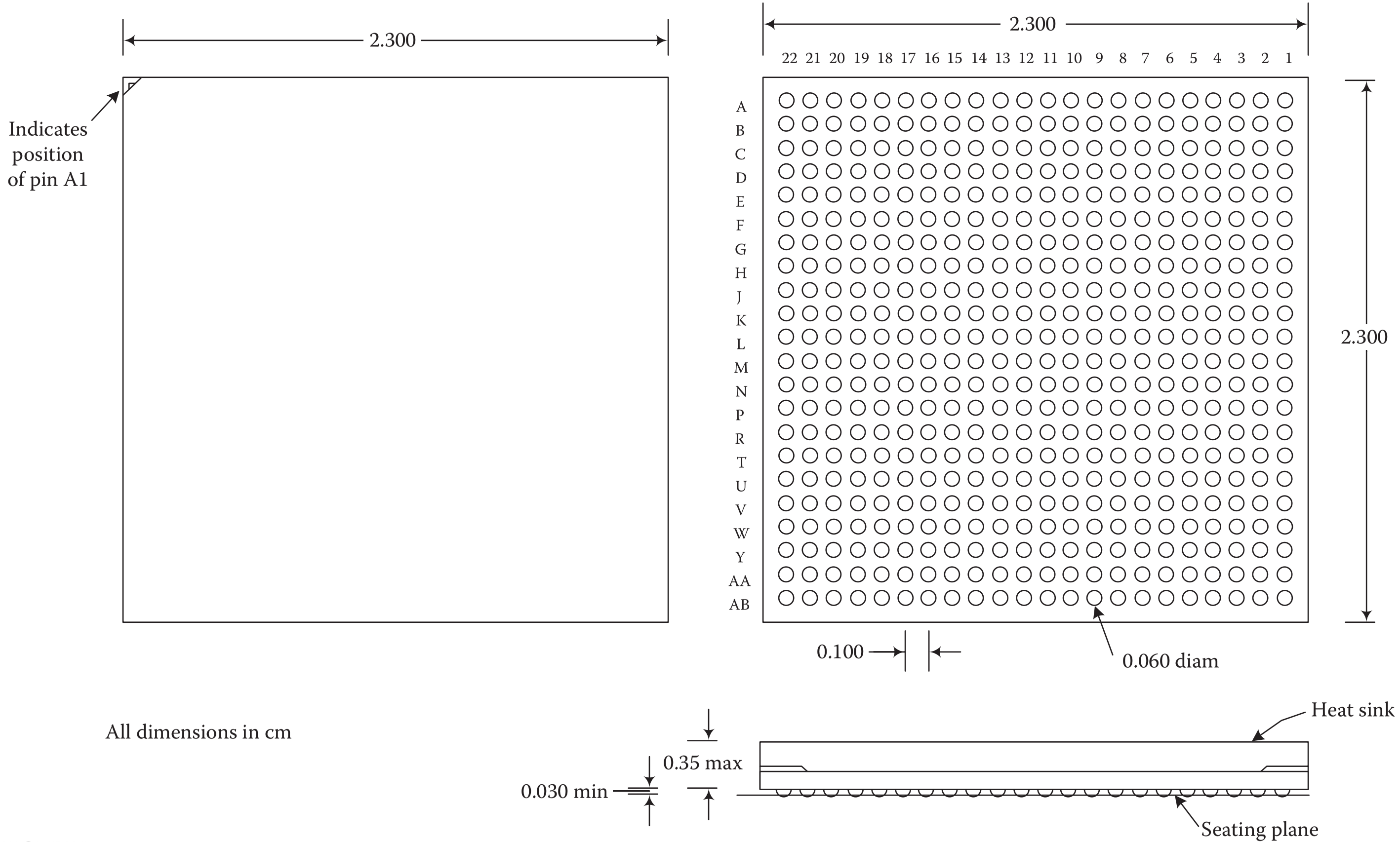


FIGURE L.3
SOIC package with 20 pins.



All dimensions in cm

FIGURE L.4
BGA package with 484 pins and an integrated heat sink.

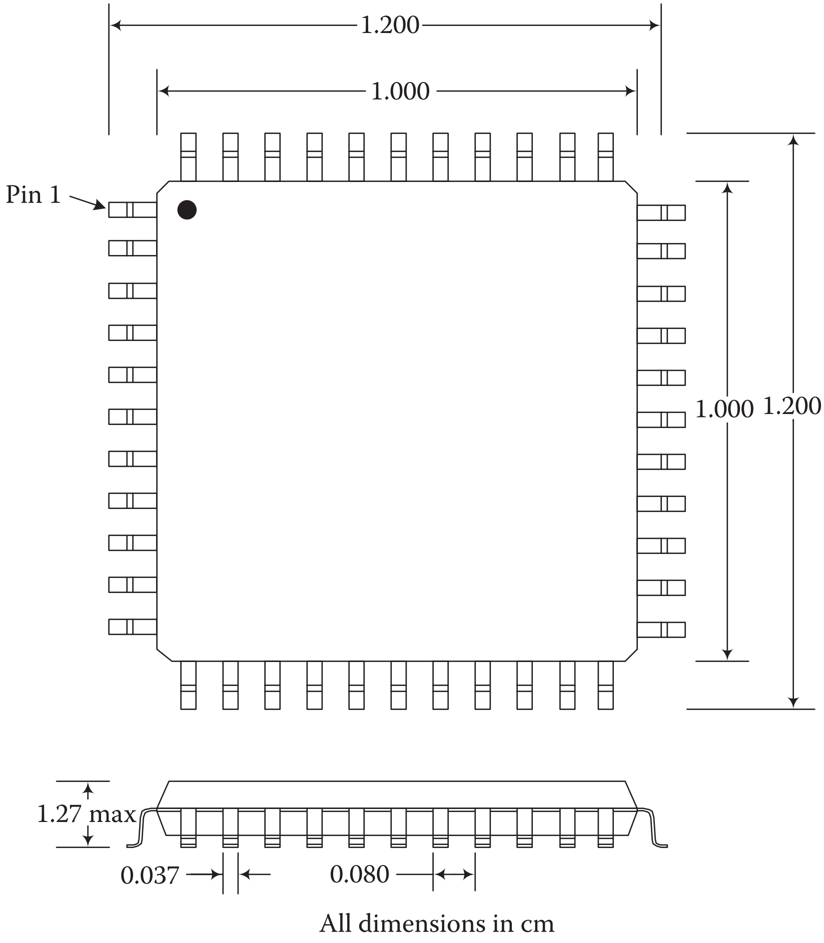


FIGURE L.5
Plastic quad flat pack package with 44 pins.

involves the bonding of 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 the use of bare die mounted on polyamide film with metal traces on it (tape automated bonding). An example of this approach is chip on flex.

L.2.5 Multichip Modules

Multichip modules (MCMs) [27–34] may use either THT or surface mount technology; the feature that distinguishes them is the placement of more

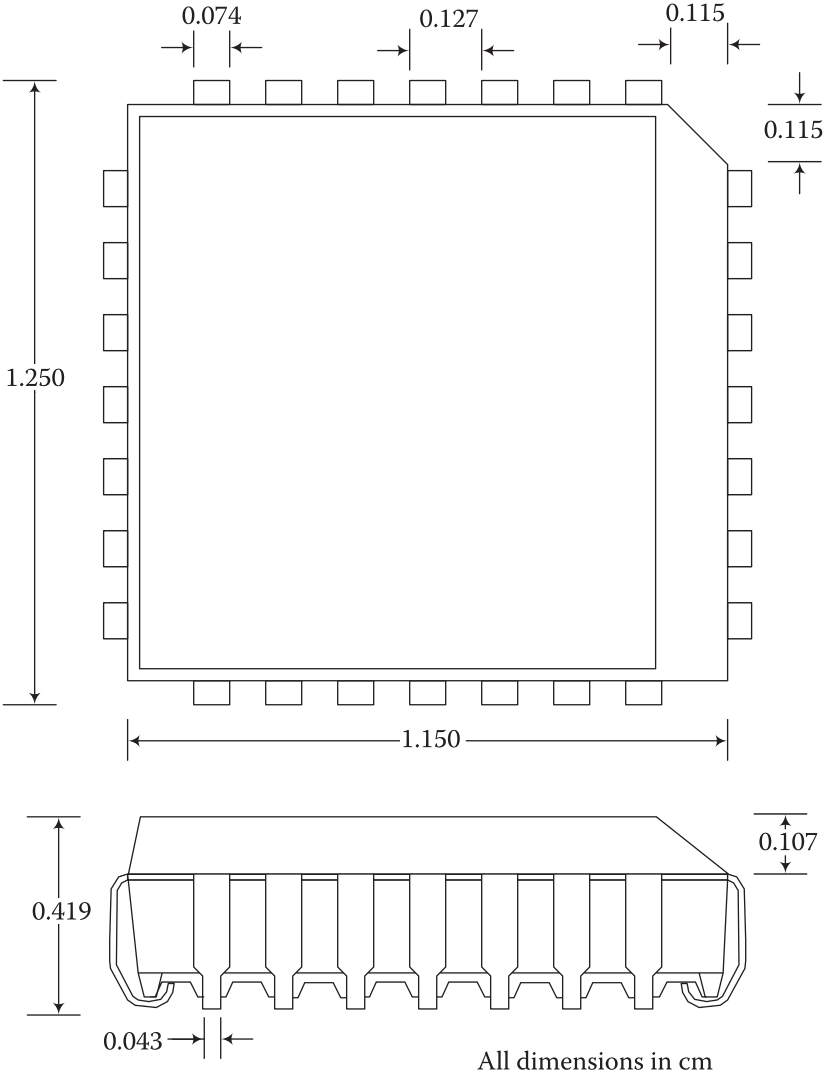


FIGURE L.6
Plastic J-lead chip carrier with 28 pins.

than one die in a single package. Some modules use die arranged in a single plane, whereas others stack the chips vertically to significantly reduce the package footprint. An example of the latter approach is the memory cube, in which DRAMs are stacked vertically. In either case, the board area consumed is significantly less than if the die were all packaged individually.

MCMs have been developed with a number of material technologies, each having its own cost-performance tradeoff. MCM-C technology uses ceramic

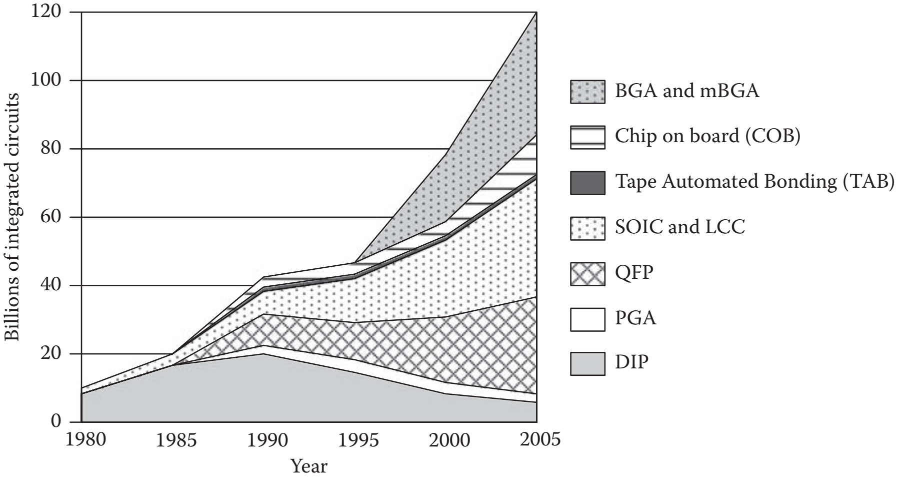


FIGURE L.7
Integrated circuit package usage by type. (Based on data from Altera Corporation, http://www.altera.com.)

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 uses laminated organic layers such as polyamide for reduced cost.

L.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 THTs such as DIPs and PGAs. On the other hand, the overall market for integrated circuits has grown explosively. This has lead to the increased use of conventional surface mount technology, BGAs, and bare die (chip on board). These trends can be seen in Figure L.7.

L.3 General Considerations

There are several general requirements for any integrated circuit package:

- It must provide an adequate number of electrical connections to the outside world (usually called pins) without imposing long signal delays.

- It must be able to conduct heat sufficiently from the operating circuit.
- It must be able to withstand elevated temperatures imposed by the circuit operation.
- It must be able to withstand the thermal cycling associated with normal circuit operation, without imposing mechanical failure attributable to thermal stresses.
- It must be able to protect the integrated circuit from the chemical environment, especially moisture and ionic contaminants.
- It must be able to protect the circuit from mechanical vibration, mechanical shock, and stresses.
- It must be capable of easy handling, testing, and assembly into systems.

Broadly, these requirements can be categorized as electrical, thermal, chemical, and mechanical.

L.3.1 Electrical Considerations

An integrated circuit package must provide the required electrical connections without imposing undue signal delays attributable to parasitic inductances, capacitances, or resistances in the leads.

VLSI circuits often require pin counts in the hundreds, and this requirement is increasing steadily. This is illustrated for the case of ASICs in Figure L.8 [5]. Therefore, the areal pin density is a commonly used figure of merit for packages. Packages having leads arranged in straight rows along two sides (such as DIPs) have low pin density. On the other hand, packages using 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 while often providing 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

$$t_{flight} = \frac{l}{c_0 / \sqrt{\varepsilon_r \mu_r}}, \tag{L.1}$$

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 for the medium. If the circuit propagation delay is less than the time

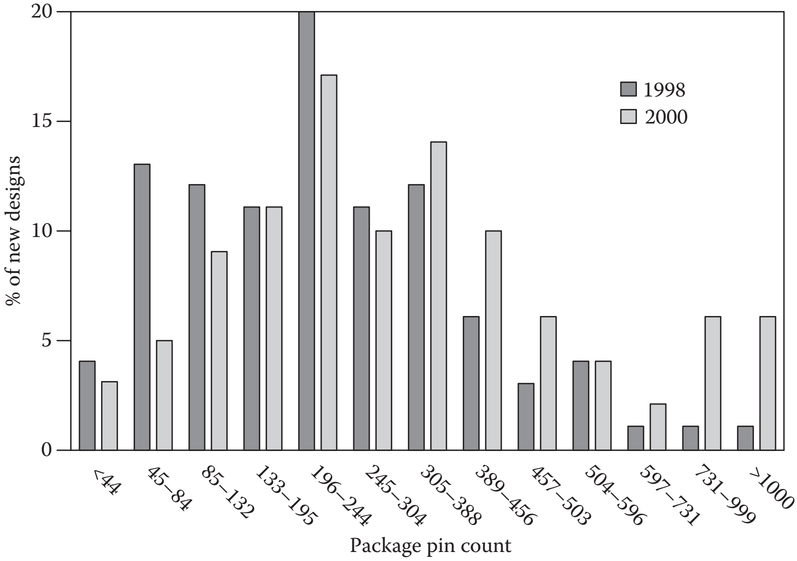


FIGURE L.8
Percentage of ASIC starts versus pin count (Based on data from Altera Corporation, http://www.altera.com.)

of flight, then the transmission model should be used. Otherwise, a lumped element model is applicable. (For example, the time of flight for an electrical signal traveling along a circuit trace on a board made of FR-4 is 17 ps/cm.) In practice, the lumped element model can often be used for traces on circuit boards, whereas the transmission line model must be used for network connections.

Figure L.9 illustrates the use of a lumped element model for the case in which a packaged circuit is driving the input to another packaged circuit. Table L.1 provides some typical values of the parasitics associated with

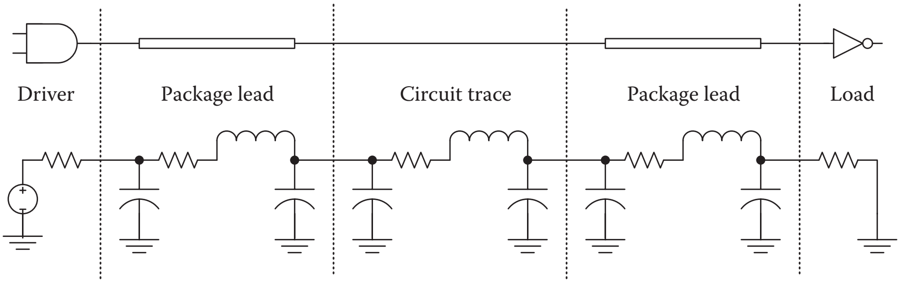


FIGURE L.9
Lumped element model for a packaged circuit driving another packaged circuit.

TABLE L.1

Typical Parasitics and Signal Delays Associated with Two Different Package Approaches

	W/B w/ PGA	F/C w/ BGA
Inductance	10 nH	1.5 nH
Capacitance	12 pF	4 pF
Resistance	20 Ω	2 Ω
Lead signal delay	700 ps	100 ps

W/B, Wafer bonded; F/C, flip chip.
Source: Based on Blackwell, G. R., *The Electronic Packaging Handbook*. CRC Press (in cooperation with IEEE Press), Boca Raton, FL, 2000.

package leads [1]. These numbers, although 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 both the dielectric constant and the loss tangent. The power dissipation, and development of heat, in the insulator is directly proportional to the loss tangent (also referred to as the dissipation factor). 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, also has a similar dielectric constant but is relatively lossy. These properties of packaging insulators are summarized in Table L.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 L.3, copper is superior in this regard and finds use in substrate conductors. Aluminum is used almost exclusively for bonding pads, whereas both gold and aluminum have been used for wire bonds.

TABLE L.2

Relative Permittivities (dielectric constants) and Loss Tangents of Insulating Materials Used in Digital Integrated Circuit Packages

Material	Dielectric constant ε_r @1MHz	Loss tangent $(\times 10^4)$@25$^\circ C$, 1MHz
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: Based on M. G. Pecht et al. *Electronic Packaging Materials and Their Properties*, CRC Press, Boca Raton, FL, 1999.

TABLE L.3

Electrical Resistivities of Conductors Commonly Used in Integrated Circuit Packages

Metal	$\rho\,(\mu\Omega cm)$
Copper	1.7
Gold	2.2
Aluminum	2.65

TABLE L.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% (Al2O3)	30
Metals	
Silver (Ag)	428
Copper (Cu)	397
Gold (Au)	317
Aluminum (Al)	230
Nickel (Ni)	88

Source: Based on M. G. Pecht et al. *Electronic Packaging Materials and Their Properties*, CRC Press, Boca Raton, FL 1999.

L.3.2 Thermal Considerations

The important thermal considerations are heat dissipation [36–38] and thermal expansion. Efficient heat removal is necessary 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 the circuit lifetimes decrease strongly with operating temperature. Thermal expansion must be considered because the integrated circuit package uses 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:

$$q = -k\nabla T \,, \tag{L.2}$$

where q is the heat flow in watts per square centimeter, k is the thermal conductivity of the solid in watts per centimeter per kelvin, and ∇T is the three-dimensional temperature gradient in kelvins per centimeter. 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 having a cross-sectional area of A, a thickness of l, and a thermal conductivity of k, the thermal resistance is given by

$$\theta = \frac{kl}{A} \,. \tag{L.3}$$

The one-dimensional heat flow is given by

$$Q = \frac{\Delta T}{\theta} \,, \tag{L.4}$$

where Q is the heat flow in watts (analogous to electrical current), ΔT is the temperature difference in kelvins (analogous to potential difference), and θ is the thermal resistance in watts per kelvin (analogous to electrical resistance).

For a dissipating integrated circuit, the junction temperatures can be calculated from

$$T_j = T_a + P_d \theta_{ja} \,, \tag{L.5}$$

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,

$$\theta_{ja} = \theta_1 + \theta_2 + \theta_3 + \ldots \tag{L.6}$$

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 L.4. Silicon has three times the thermal conductivity of gallium arsenide (GaAs); therefore, GaAs integrated circuits often require mechanical thinning to achieve efficient heat transfer. Diamond exhibits superior thermal conductivity compared with 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 a 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 the thermal resistance. For example, the 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 with the use of 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.

L.3.3 Chemical Considerations

An integrated circuit package must protect the circuit from its chemical environment, both during storage and operation. Also the many materials used in its manufacture must themselves 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 take up appreciable amounts of water from the air. If these parts are not baked out adequately before assembly, the sudden temperature rise associated with soldering will cause package failure (*"popcorning"*) [39–43] attributable to the rapid vaporization of the water. During operation of assembled systems, the 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 the use of an 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% iron, 29% nickel, and 17% cobalt). Here, the Kovar package and lid are hermetically sealed using a glass frit. The thermal expansion coefficient of Kovar (5.1–5.9 ppm/K) closely matches thermal expansion coefficients of commonly used sealing glasses (5.25–6.96 ppm/K), therefore minimizing thermal stresses associated with the seal. However, the thermal conductivity of Kovar is relatively poor (15.5–17 W/mK) so that copper alloys are used for high-power hermetic packages.

L.3.4 Mechanical Considerations

Generally speaking, mechanical failure mechanisms in an integrated circuit package may be classified as *instantaneous mechanical overload* or *progressive* in nature [44]. Instantaneous overloading problems include ductile deformation and brittle fracture. Progressive failure mechanisms include fatigue crack growth and creep deformation.

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:

$$\sigma = \frac{F}{A}. \tag{L.7}$$

If the applied stress exceeds the yield stress σ_y of the metal, permanent deformation will occur. This may result in a broken electrical connection. In practice, packaging engineers must ensure that the yield stress is never exceeded. The yield stresses of metals commonly used in integrated packages are given in Table L.5. 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

$$\sigma = \frac{YK_{Ic}}{\sqrt{a}}, \tag{L.8}$$

where K_{Ic} 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 L.6. It can be seen that beryllia, alumina, and silicon carbide are superior in this regard.

TABLE L.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

TABLE L.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

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, there is a gradual increase in plastic deformation over a long period of time, which gives 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 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

$$L_o = B\left(\Delta\sigma\right)^{-\theta_b}, \tag{L.9}$$

where $\Delta\sigma$ is the peak-to-peak amplitude of the time-varying stress. 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 periodic stress. Fatigue is an important failure mechanism for solder bumps used in flip chip technology and for wire bonds in plastic packages.

L.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. Either 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 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.

L.4.1 Wire Bond Process

The wire bonding approach is commonly used in conjunction with both plastic and ceramic packages. By this process, electrical connections between the die and the pins are made using fine gold or aluminum wires. The process flow for a wire bond process is outlined in Figure L.10.

After 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 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 are packaged, resulting in considerable cost savings.

Die bonding involves the attachment of known good die to a ceramic substrate or a metal lead frame. Ceramic substrates are commonly alumina-silica mixtures (90-99% Al_2O_3, balance SiO_2) or beryllia, but many other materials are available. Metal lead frames are made from a copper alloy or Kovar (a metallic alloy of 54% iron, 29% nickel, and 17% cobalt). Electrical connections are made from the die to the package leads by wire bonding.

There are several variations of the wire bonding process. One method in common use is the thermosonic ball-wedge technique [45] illustrated in Figure L.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. In either case, the end result is 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^{o}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. After 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.

* Cleavage is the separation of the crystal along natural crystal planes, called "cleavage planes." For example, silicon crystals cleave on [110] planes. For the case of a silicon (001) wafer, cleavage on these planes results is rectangular die, with edges oriented by 54.7^o to the top surface.

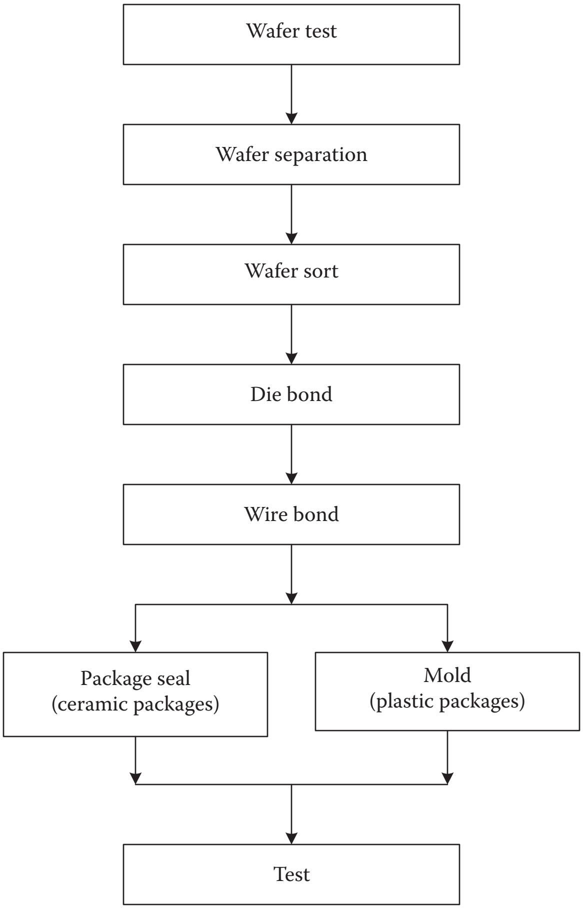


FIGURE L.10
Process flow for a wire bond packaging process.

This process is repeated until all connections have been made. After 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

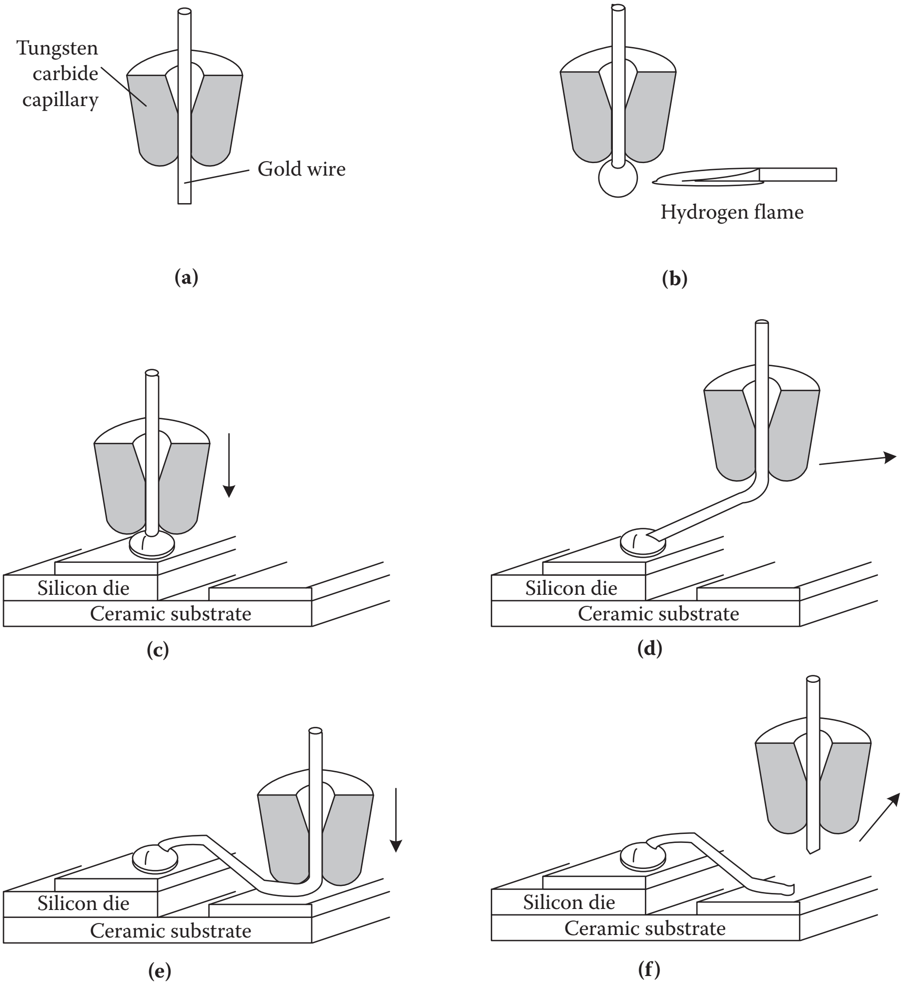


FIGURE L.11
Wire bonding using the thermosonic ball-wedge approach. (a) A gold wire, 10–50 μm in diameter, is drawn through a tungsten carbide capillary. (b) A hydrogen microtorch or a capacitive discharge is used to form a gold ball on the end of the wire. (c) The gold ball is bonded to an aluminum pad on the heated (~150°C) silicon die using a vertically applied force and ultrasonic vibration (~50 kHz). (d) The capillary tip is pulled over to metal pad on the heated substrate, and (e) 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. (Based on Ghandhi, S.K., *VLSI Fabrication Principles*, 2nd Ed., Wiley, New York, 1994.)

bond wires before 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.

L.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 lead-tin eutectic or a lead-indium alloy. Figure L.12 illustrates the implementation of a lead-tin solder bump over an aluminum pad.

The process flow for the flip chip approach is outlined in Figure L.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 lead/tin eutectic) that forms an excellent electrical and mechanical connection between the flip chip and the package. The surface tension of the molten solder ensures proper alignment between the flip chip and the package. After reflow, the package is sealed. Finally, the packaged circuit is tested.

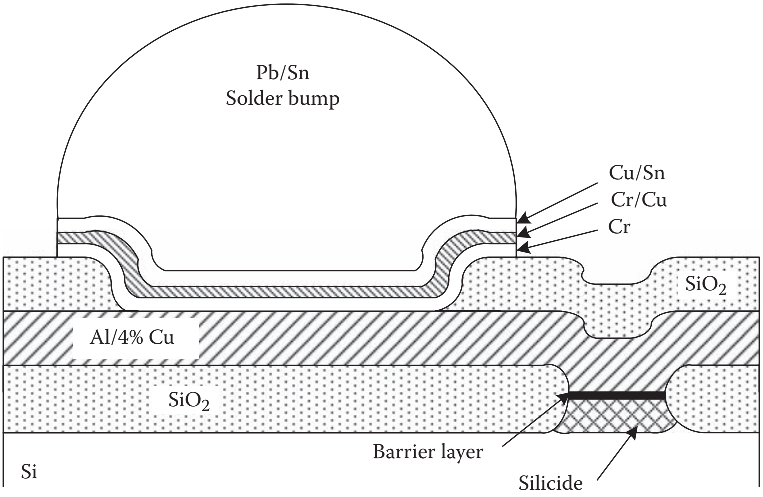


FIGURE L.12
A solder bump on a silicon wafer.

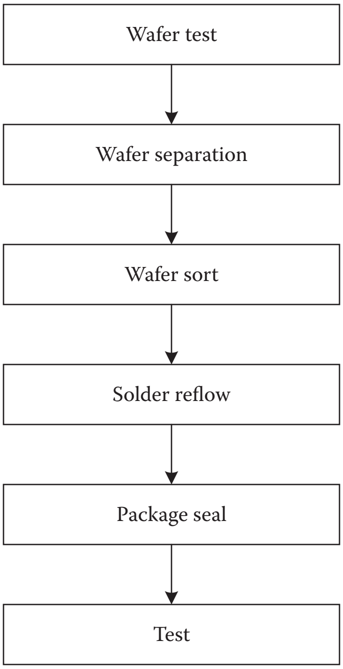


FIGURE L.13
Process flow for a flip chip packaging process.

Elimination of the bonding wires in "flip chip" packages allows the use of bonding pads that cover the entire chip area rather than just the periphery. In addition to greater pin density and improved heat removal, this avoids the signal delays associated with the inductance and resistance of the bonding wires.

A problem encountered in flip chip packages is the thermal fatigue of the solder bump connections. It is found that the use of epoxy underfill with the solder bumps greatly enhances the fatigue lifetimes of solder bump connections. This is evident from the results compiled in Table L.7 for common solder bump alloys.

L.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 the efficient removal of

TABLE L.7

Thermal Fatigue Lifetimes for Solder Bump Alloys (–40 to +125°C cycles), Normalized to the Lifetime for Eutectic Tin-Lead Solder (63% tin/37% lead) with No Underfill

Bump alloy	T(reflow) (°C)	Normalized life (no underfill)	Normalized life (w/ epoxy underfill)
63 Sn/37 Pb	230	1.0	15
50 In/50 Pb	260	2–3	>30
37 In/63 Pb	290	2–3	>30
3.5 Ag/96.5 Sn	260	0.5	11
5 Sb/95 Sn	280	0.3	11
Sn/Pb/Cd/In	230	1.0	13
Sn/Ag/Cu/Sb	260	1.0	13

large quantities of heat (~10^2 W). Moreover, these packages must be compact, lightweight, inexpensive, and reliable.

There are five basic types of integrated circuit packages: through-hole packages, surface mount packages, chip-scale packages, bare die, and module assemblies. THT packages have metal pins that may be inserted through holes drilled in the circuit board for soldering. Surface mount technology packages use metal leads that can be soldered to a single surface of the printed circuit board. They are much smaller and lighter weight than through-hole packages, for a given number of electrical connections. In addition, they are more resistant to mechanical shock compared with 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 with bare die. Usually, chip-scale 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 use.

The process of packaging an integrated circuit involves either 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.

References

1. Blackwell, G.R., *The Electronic Packaging Handbook*. CRC Press in cooperation with IEEE Press, Boca Raton, FL, 2000.
2. Joint Electron Device Engineering Council, http://www.jedec.org.
3. Institute of Electrical and Electronics Engineers, http://www.ieee.org.
4. Intel Corporation, http://www.intel.com.
5. Altera Corporation, http://www.altera.com.
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. *Semiconduct. 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 Electronics Packaging 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 Electronic 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 Manufact. Technol.*, 6, 298, 1983.
11. Mattei, C., and Agrawal, A.P., Electrical characterization of BGA packages. *Proc. 47th Electronic 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 Japan Int. Electronic Manufact. Technol. Symp.,* 41, 1993.
14. Rao, S.T., Ball grid array assembly issues in manufacturing. *Proc. 16th IEEE/CPMT Int. Electronics Manufact. Technol. Symp.*, 347, 1994.
15. Thompson, P., Chip-scale packaging. *IEEE Spectrum,* 34, 36, 1997.
16. Okuno, A., Fujita, N., and Ishikana, Y., Low cost and high reliability extremity CSP packaging technology. *Proc. 49th Electronic Components Technol. Conf.*, 1201, 1999.
17. Elenius, P., The Ultra CSP™ wafer scale package. *Proc. 2nd Electronics Packaging Technol. Conf.*, 83, 1998.
18. Arnold, R., Chip scale package versus direct chip attach (CSP vs. DCA). *Proc. 50th Electronic 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, http://www.intel.com, 1999.
21. Rochat, G., COB and COC for low cost and high density package. *Proc. 17th IEEE/CPMT Int. Electronics Manufact. 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 Electronic 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 Electronic Components Technol. Conf.*, 638, 1997.
25. O'Malley, G., Giesler, J., Machuga, S., The importance of material selection for flip chip on board assemblies. *Proc. 44th Electronic Components Technol. Conf.*, 387, 1994.
26. *Understanding the Quality and Reliability Requirements for Bare Die Applications*, Micron Technology, Inc. Technical Note, http://www.micron.com, 2002.
27. Charles, H.K., Packaging with multichip modules. *Proc. 13th IEEE/CHMT Electronics Manufact. Technol. Symp.*, 206, 1992.
28. Vasquez, B., and Tippins, F., Multichip modules: packaging solutions for size performance integration. *Int. Integrated Reliability Workshop Final Report*, 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. Electrical Performance of Electronic Packaging*, 132, 1998.
31. Iqbal, A., Swaminathan, M., Nealon, M., and Omer, A., Design tradeoffs 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 & 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. M. G. Pecht, R. Agarwal, P. McCluskey, T. Dishongh, S. Javadpour, and R. Mahajan, *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 DIP's. *IEEE Trans. Components Hybrids Manufact. Technol.*, 4, 455, 1981.
37. Mulgaonker, S., Chambers, B., and Mahalingam, M., An assessment of the thermal performance of the PBGA family. *11th IEEE Semiconductor Thermal Measure. Manage. Symp.,* 17, 1995.
38. Edwards, D., Hwang, M., and Stearns, B., Thermal enhancement of IC packages, Proc. 10th IEEE/CPMT Semiconductor Thermal Measurement and Management 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 Packaging Manufact. Technol., Part B Adv. Packaging,* 18, 491, 1995.
41. Gannamani, R., and Pecht, M., An experimental study of popcorning in plastic encapsulated microcircuits. *IEEE Trans. Components Packaging Manufactur. 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 Packaging 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 Packaging Technol.*, 25, 56, 2002.
44. Pecht, M., *Integrated Circuit, Hybrid and Multichip Module Package Design Guidelines: A Focus on Reliability*. Wiley, New York, 1994.
45. Ghandhi, S.K., *VLSI Fabrication Principles,* 2nd Ed., Wiley, New York, 1994.

Index

ELECTRICAL ENGINEERING

SECOND EDITION

Digital Integrated CIRCUITS

Analysis and Design

Exponential improvement in functionality and performance of digital integrated circuits has revolutionized the way we live and work. The continued scaling down of MOS transistors has broadened the scope of use for circuit technology to the point that texts on the topic are generally lacking after a few years.

The second edition of **Digital Integrated Circuits: Analysis and Design** focuses on timeless principles with a modern interdisciplinary view that will serve integrated circuits engineers from all disciplines for years to come. Providing a revised instructional reference for engineers involved with Very Large Scale Integrated Circuit design and fabrication, this book delves into the dramatic advances in the field, including new applications and changes in the physics of operation made possible by relentless miniaturization.

This book was conceived in the versatile spirit of the field to bridge a void that had existed between books on transistor electronics and those covering VLSI design and fabrication as a separate topic. Like the first edition, this volume is a crucial link for integrated circuit engineers and those studying the field, supplying the cross-disciplinary connections they require for guidance in more advanced work.

For pedagogical reasons, the author uses SPICE level 1 computer simulation models but introduces BSIM models that are indispensable for VLSI design. This enables users to develop a strong and intuitive sense of device and circuit design by drawing direct connections between the hand analysis and the SPICE models.

With four new chapters, more than 200 new illustrations, numerous worked examples, case studies, and support provided on a dynamic Web site, this text significantly expands concepts presented in the first edition.

CRC Press
Taylor & Francis Group
an informa business

6000 Broken Sound Parkway, NW
Suite 300, Boca Raton, FL 33487
270 Madison Avenue
New York, NY 10016
2 Park Square, Milton Park
Abingdon, Oxon OX14 4RN, UK

6987X
ISBN: 978-1-4200-6987-7
90000
9 781420 069877
www.crcpress.com