

FOURTH EDITION

MICROELECTRONIC CIRCUITS

Adel S. Sedra

University of Toronto

Kenneth C. Smith

*University of Toronto and Hong Kong
University of Science and Technology*

New York Oxford

OXFORD UNIVERSITY PRESS

1998

Oxford University Press

Oxford New York
Athens Auckland Bangkok Bogota Bombay Buenos Aires
Calcutta Cape Town Dar es Salaam Delhi Florence Hong Kong
Istanbul Karachi Kuala Lumpur Madras Madrid Melbourne
Mexico City Nairobi Paris Singapore Taipei Tokyo Toronto Warsaw

and associated companies in

Berlin Ibadan

Copyright © 1998, 1991, 1987, 1982 by Oxford University Press, Inc.

Published by Oxford University Press, Inc.,
196 Madison Avenue, New York, New York, 10016
<http://www.oup-usa.org>

Oxford is a registered trademark of Oxford University Press

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system,
or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording,
or otherwise, without the prior permission of Oxford University Press.

Staff

Senior Acquiring Editor: *Bill Zobrist*
Editorial/Production/Design Manager: *Elyse Dubin*
Assistant Acquiring Editor: *Krysia Bebeck*
Production Coordinator: *John Butco*
Editorial Assistant: *Jasmine Urmeneta*
Composition: *York Graphic Services*
Production Services: *York Production Services*
Director of Marketing: *Scott Burns*
Assistant Sales Manager: *Marcy Levine*
Advertising Manager: *David Walker*
Copywriter: *Leigh Ann Florek*
Cover Design: *Ed Berg Designs*
Cover Art Direction: *Linda Roppolo*
Manufacturing Director: *Ben Lee*
Manufacturing Controller: *Robin Meredith-Clement*

Library of Congress Cataloging-in-Publication Data

Sedra, Adel S.
Microelectronic circuits / Adel S. Sedra, Kenneth C. Smith. — 4th ed.
p. cm. — (Oxford series in electrical and computer engineering)
Includes bibliographical references and index.
ISBN 0-19-511663-1
1. Electronic circuits. 2. Integrated circuits. I. Smith,
Kenneth Carless. II. Title. III. Series.
TK7867 S34 1997
621.381—dc22

97-11254
CIP

Printed in the United States of America on acid-free paper

Cover illustration: The chip shown is the ADXL-50 surface-micromachined accelerometer. For the first time, sensor and signal conditioning are combined on a single monolithic chip. In its earliest application, it was a key factor in the improved reliability and reduced cost of modern automotive airbag systems. Photo reprinted with permission of Analog Devices, Inc.

PREREQ

ORGAN

CHAPTER 3

Diodes

	Introduction		3.7 Rectifier Circuits
3.1	The Ideal Diode		3.8 Limiting and Clamping Circuits
3.2	Terminal Characteristics of Junction Diodes		3.9 Special Diode Types
3.3	Physical Operation of Diodes		3.10 The SPICE Diode Model and Simulation Examples
3.4	Analysis of Diode Circuits		Summary
3.5	The Small-Signal Model and Its Application		Bibliography
3.6	Operation in the Reverse Breakdown Region—Zener Diodes		Problems

INTRODUCTION

In the previous chapter we dealt almost entirely with linear circuits; any nonlinearity, such as that introduced by amplifier output saturation, was considered a problem to be solved by the circuit designer. However, there are many other signal-processing functions that can be implemented only by nonlinear circuits. Examples include the generation of dc voltages from the ac power supply and the generation of signals of various waveforms (e.g., sinusoids, square waves, pulses, etc). Also, digital logic and memory circuits constitute a special class of nonlinear circuits.

The simplest and most fundamental nonlinear circuit element is the diode. Just like a resistor, the diode has two terminals; but unlike the resistor, which has a linear (straight-line) relationship between the current flowing through it and the voltage appearing across it, the diode has a nonlinear $i-v$ characteristic.

This chapter is concerned with the study of diodes. In order to understand the essence of the diode function, we begin with a fictitious element, the ideal diode. We then introduce the silicon junction diode, explain its terminal characteristics, and provide techniques for

the analysis of diode circuits. The latter task involves the important subject of device modeling.

To understand the origin of the diode terminal characteristics, we consider its physical operation. Our study of the physical operation of the pn junction and of the basic concepts of semiconductor physics is intended to provide a foundation for understanding not only the characteristics of junction diodes but also those of the bipolar junction transistor, studied in the next chapter, and field-effect transistors, studied in Chapter 5.

Of the many applications of diodes, their use in the design of rectifiers (which convert ac to dc) is the most common. Therefore, we shall study rectifier circuits in some detail and briefly look at a number of other diode applications. Further nonlinear circuits that utilize diodes and other devices will be found throughout the book and in particular in Chapter 12.

Although most of this chapter is concerned with the study of silicon pn -junction diodes, we briefly consider some specialized diode types, including the photodiode and the light-emitting diode. The chapter concludes with a description of the diode model utilized in the SPICE circuit-simulation program. We also present a design example that illustrates the use of SPICE simulation.

3.1 THE IDEAL DIODE

The ideal diode may be considered the most fundamental nonlinear circuit element. It is a two-terminal device having the circuit symbol of Fig. 3.1(a) and the i - v characteristic shown

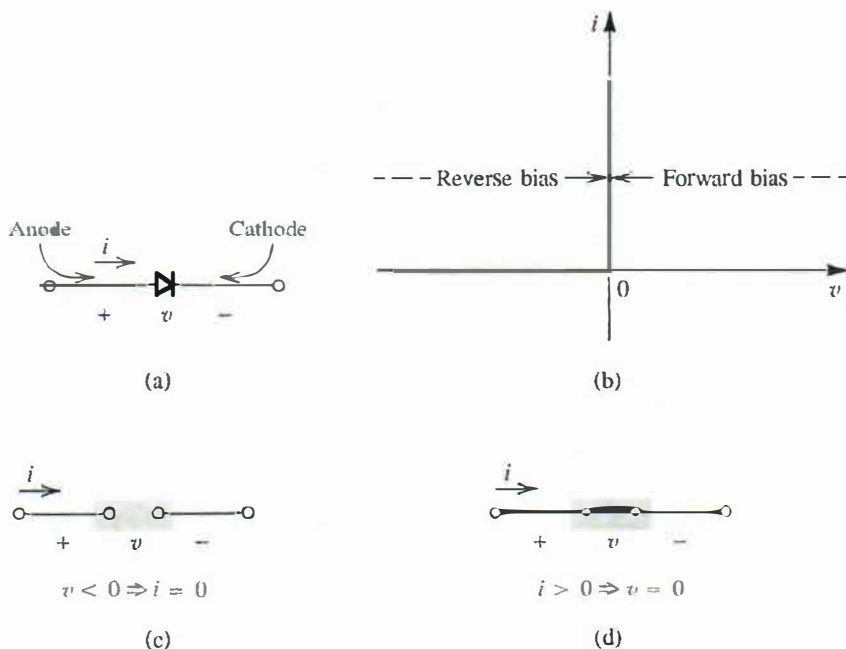


Fig. 3.1 The ideal diode: (a) diode circuit symbol; (b) i - v characteristic; (c) equivalent circuit in the reverse direction; (d) equivalent circuit in the forward direction.

in Fig. 3.1(b). The terminal characteristic of the ideal diode can be interpreted as follows: If a negative voltage [relative to the reference direction indicated in Fig. 3.1(a)] is applied to the diode, no current flows and the diode behaves as an open circuit [Fig. 3.1(c)]. Diodes operated in this mode are said to be **reverse-biased**, or operated in the reverse direction. An ideal diode has zero current when operated in the reverse direction and is said to be **cut off**.

On the other hand, if a positive current [relative to the reference direction indicated in Fig. 3.1(a)] is applied to the ideal diode, zero voltage drop appears across the diode. In other words, the ideal diode behaves as a short circuit in the *forward* direction [Fig. 3.1(d)]; it passes any current with zero voltage drop. A forward-conducting diode is said to be **turned on**, or simply, **on**.

From the above description it should be noted that the external circuit must be designed so as to limit the forward current through a conducting diode, and the reverse voltage across a cutoff diode, to predetermined values. Figure 3.2 shows two diode circuits that illustrate this point. In the circuit of Fig. 3.2(a) the diode is obviously conducting. Thus its voltage drop will be zero, and the current through it will be determined by the $+10\text{-V}$ supply and the $1\text{-k}\Omega$ resistor as 10 mA . The diode in the circuit of Fig. 3.2(b) is obviously cut off, and thus its current will be zero, which in turn means that the entire 10-V supply will appear as reverse bias across the diode.

The positive terminal of the diode is called the **anode** and the negative terminal the **cathode**, a carryover from the days of vacuum-tube diodes. The i - v characteristic of the ideal diode (conducting in one direction and not in the other) should explain the choice of its arrowlike circuit symbol.

As should be evident from the preceding description, the i - v characteristic of the ideal diode is highly nonlinear; it consists of two straight-line segments at 90° to one another. A nonlinear curve that consists of straight-line segments is said to be **piecewise linear**. If a device having a piecewise-linear characteristic is used in a particular application in such a way that the signal across its terminals swings only along one of the linear segments, then the device can be considered a linear circuit element as far as that particular circuit application is concerned. On the other hand, if signals swing past one or more of the break points in the characteristic, linear analysis is no longer possible.

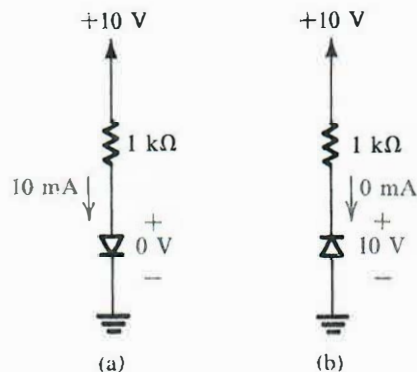


Fig. 3.2 The two modes of operation of **ideal** diodes and the use of an external circuit to limit the forward current and the reverse voltage.

A Simple Application: The Rectifier

A fundamental application of the diode, one that makes use of its severely nonlinear i - v curve, is the rectifier circuit shown in Fig. 3.3(a). The circuit consists of the series connection of a diode D and a resistor R . Let the input voltage v_I be the sinusoid shown in Fig. 3.3(b), and assume the diode to be ideal. During the positive half-cycles of the input sinusoid, the positive v_I will cause current to flow through the diode in its forward direction. It follows that the diode voltage v_D will be very small—ideally zero. Thus the circuit will have the equivalent shown in Fig. 3.3(c), and the output voltage v_O will be equal to the input voltage v_I . On the other hand, during the negative half-cycles of v_I , the diode will not conduct. Thus the circuit will have the equivalent shown in Fig. 3.3(d), and v_O will be zero. Thus, the output voltage will have the waveform shown in Fig. 3.3(e). Note that while v_I alternates in polarity and has a zero average value, v_O is unidirectional and has a finite average value or a *dc component*. Thus the circuit of Fig. 3.3(a) **rectifies** the signal and

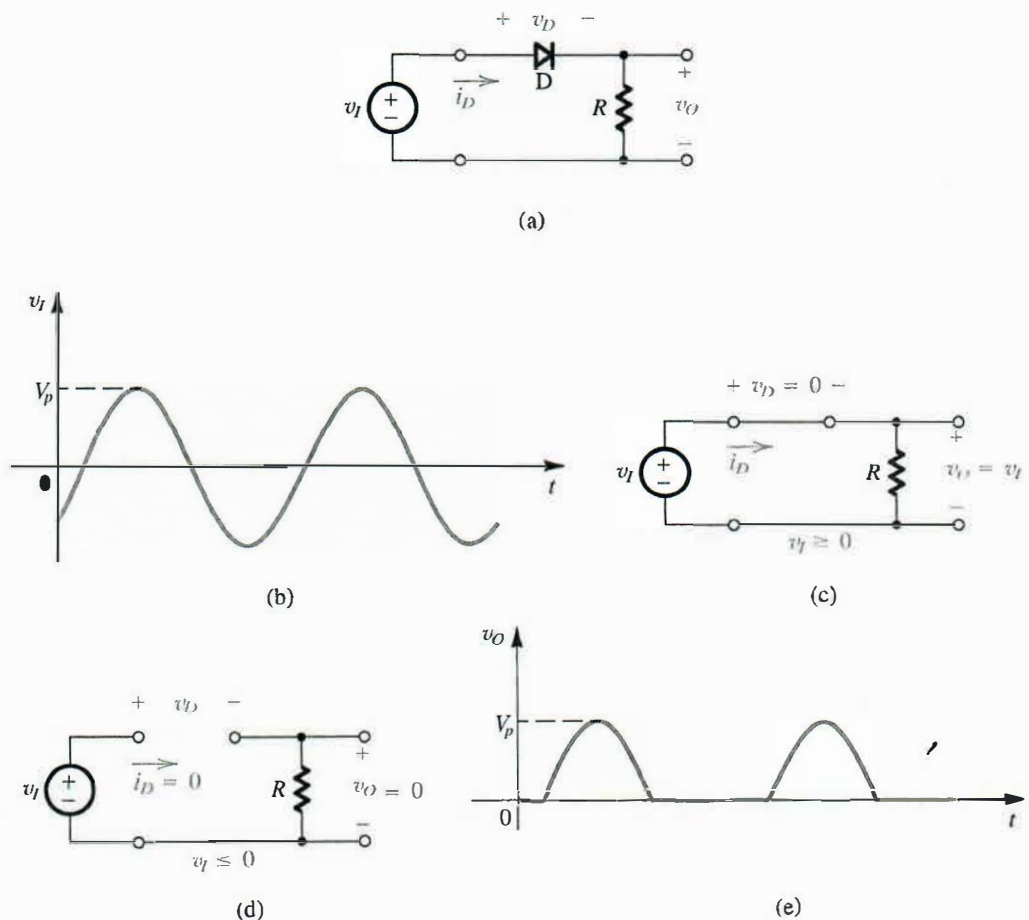


Fig. 3.3 (a) Rectifier circuit. (b) Input waveform. (c) Equivalent circuit when $v_I \geq 0$. (d) Equivalent circuit when $v_I \leq 0$. (e) Output waveform.

hence is called a **rectifier**. It can be used to generate dc from ac. We will study rectifier circuits in detail in Section 3.7.

Exercises

3.1 For the circuit in Fig. 3.3(a) sketch the transfer characteristic v_O versus v_I .

Ans. See Fig. E3.1.

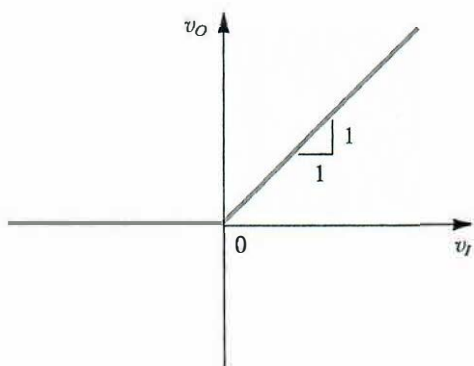


Fig. E3.1

3.2 For the circuit in Fig. 3.3(a) sketch the waveform of v_D .

Ans. See Fig. E3.2.

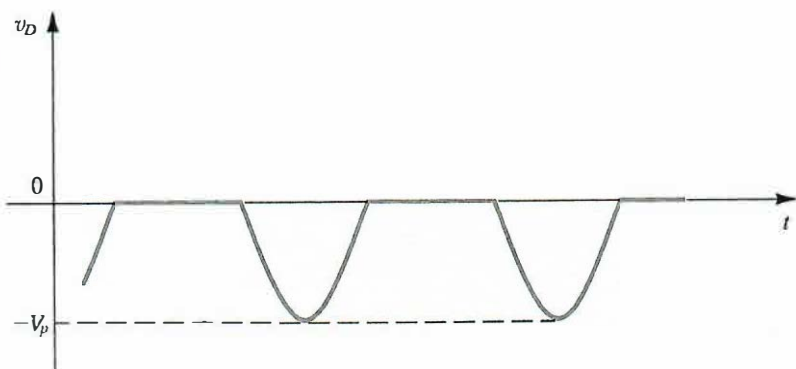


Fig. E3.2

3.3 In the circuit of Fig. 3.3(a), let v_I have a peak value of 10 V and let $R = 1 \text{ k}\Omega$. Find the peak value of i_D and the dc component of v_O .

Ans. 10 mA; 3.18 V

EXAMPLE 3.1

Figure 3.4(a) shows a circuit for charging a 12-V battery. If v_S is a sinusoid with 24-V peak amplitude, find the fraction of each cycle during which the diode conducts. Also find the peak value of the diode current and the maximum reverse-bias voltage that appears across the diode.

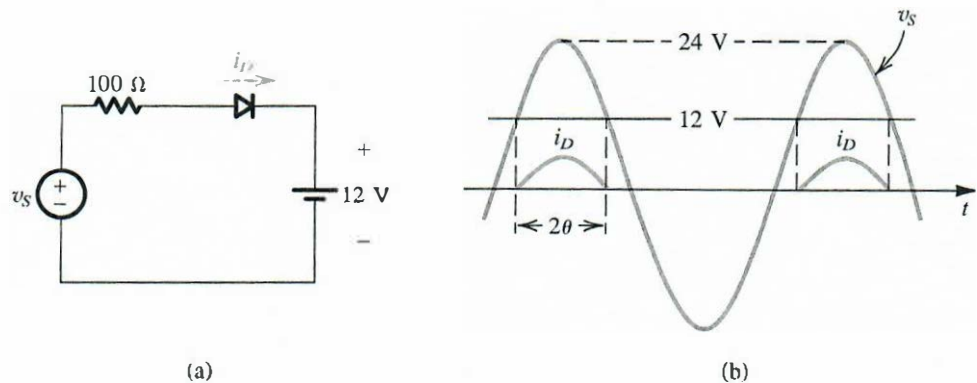


Fig. 3.4 Circuit and waveforms for Example 3.1.

SOLUTION

The diode conducts when v_S exceeds 12 V, as shown in Fig. 3.4(b). The conduction angle is 2θ , where θ is given by

$$24 \cos \theta = 12$$

Thus $\theta = 60^\circ$ and the conduction angle is 120° or one-third of a cycle.

The peak value of the diode current is given by

$$I_d = \frac{24 - 12}{100} = 0.12 \text{ A}$$

The maximum reverse voltage across the diode occurs when v_S is at its negative peak and is equal to $24 + 12 = 36 \text{ V}$.

Another Application: Diode Logic Gates

Diodes together with resistors can be used to implement digital logic functions. Figure 3.5 shows two diode logic gates. To see how these circuits function, consider a positive-logic system in which voltage values close to 0 V correspond to logic 0 (or low) and voltage values close to +5 V correspond to logic 1 (or high). The circuit in Fig. 3.5(a) has three inputs, v_A , v_B , and v_C . It is easy to see that diodes connected to +5-V inputs will conduct, thus clamping the output v_Y to a value equal to +5 V. This positive voltage at the output will keep the diodes whose inputs are low (around 0 V) cut off. Thus the output will be

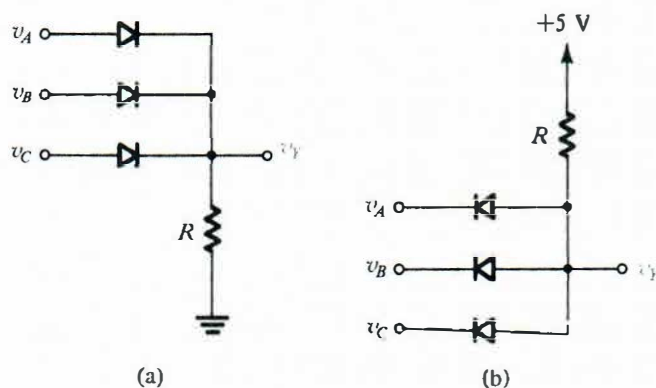


Fig. 3.5 Diode logic gates: (a) OR gate; (b) AND gate (in a positive-logic system).

high if one or more of the inputs are high. The circuit therefore implements the logic OR function, which in Boolean notation is expressed as

$$Y = A + B + C$$

Similarly, the reader is encouraged to show that using the same logic system mentioned above, the circuit of Fig. 3.5(b) implements the logic AND function,

$$Y = A \cdot B \cdot C$$

EXAMPLE 3.2

Assuming the diodes to be ideal, find the values of I and V in the circuits of Fig. 3.6.

SOLUTION

In these circuits it might not be obvious at first sight whether none, one, or both diodes are conducting. In such a case, we make a plausible assumption, proceed with the analysis, and then check whether we end up with a consistent solution. For the circuit in Fig. 3.6(a), we shall assume that both diodes are conducting. It follows that $V_B = 0$ and $V = 0$. The current through D_2 can now be determined from

$$I_{D2} = \frac{10 - 0}{10} = 1 \text{ mA}$$

Writing a node equation at B,

$$I + 1 = \frac{0 - (-10)}{5}$$

results in $I = 1 \text{ mA}$. Thus D_1 is conducting as originally assumed, and the final result is $I = 1 \text{ mA}$ and $V = 0 \text{ V}$.

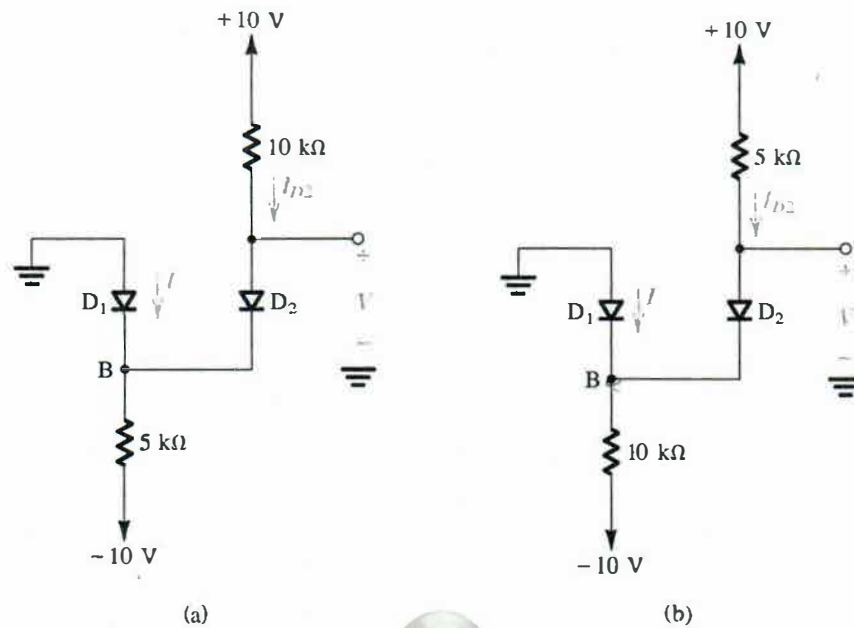


Fig. 3.6 Circuits for Example 3.2.

For the circuit in Fig. 3.6(b), if we assume that both diodes are conducting, then $V_B = 0$ and $V = 0$. The current in D_2 is obtained from

$$I_{D2} = \frac{10 - 0}{5} = 2 \text{ mA}$$

The node equation at B is

$$I + 2 = \frac{0 - (-10)}{10}$$

which yields $I = -1 \text{ mA}$. Since this is not possible, our original assumption is not correct. We start again, assuming that D_1 is off and D_2 is on. The current I_{D2} is given by

$$I_{D2} = \frac{10 - (-10)}{15} = 1.33 \text{ mA}$$

and the voltage at node B is

$$V_B = -10 + 10 \times 1.33 = +3.3 \text{ V}$$

Thus D_1 is reverse-biased as assumed, and the final result is $I = 0$ and $V = 3.3 \text{ V}$.

Exercises

3.4 Find the values of I and V in the circuits shown in Fig. E3.4.

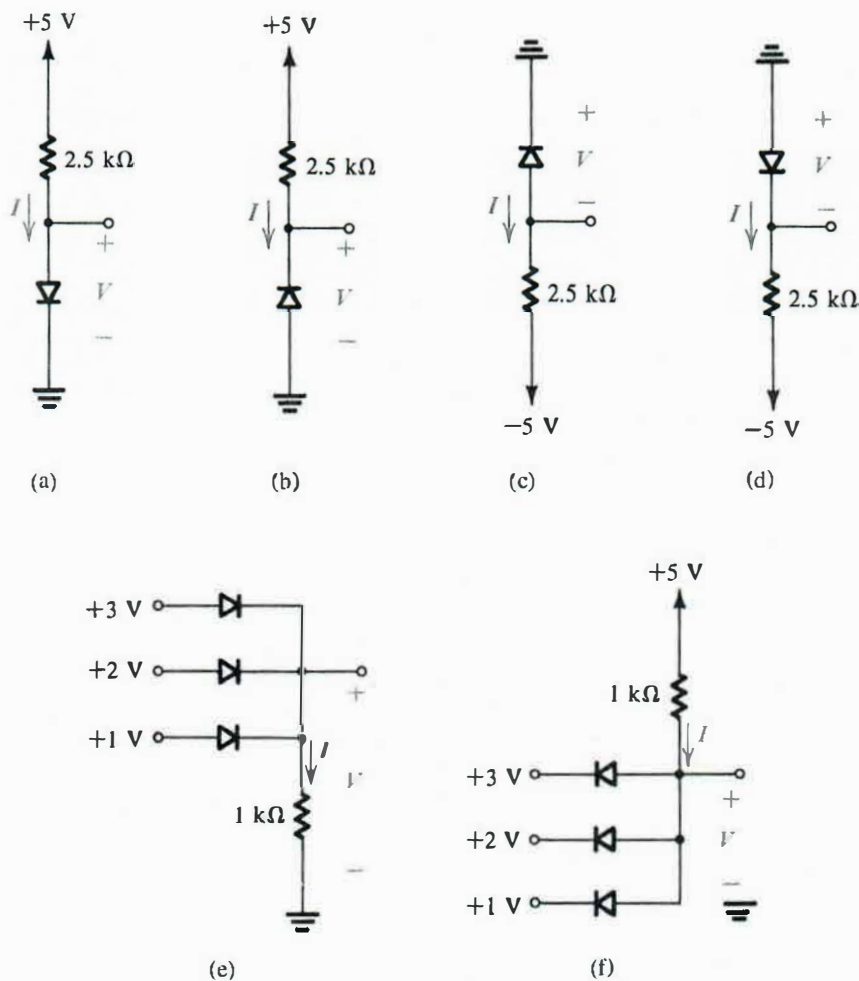
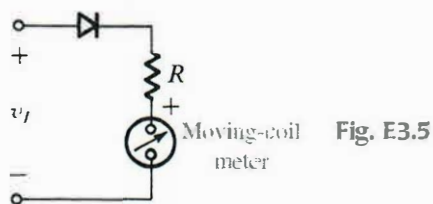


Fig. E3.4

Ans. (a) 2 mA , 0 V ; (b) 0 mA , 5 V ; (c) 0 mA , 5 V ; (d) 2 mA , 0 V ; (e) 3 mA , $+3\text{ V}$; (f) 4 mA , $+1\text{ V}$.

3.5 Figure E3.5 shows a circuit for an ac voltmeter. It utilizes a moving-coil meter that gives a full-scale reading when the *average* current flowing through it is 1 mA . The moving-coil meter has a $50\text{-}\Omega$ resistance.



Find the value of R that results in the meter indicating a full-scale reading when the input sine-wave voltage v_I is 20 V peak-to-peak. (Hint: The average value of half-sine waves is V_p/π .)

Ans. $3.133\text{ k}\Omega$

3.2 TERMINAL CHARACTERISTICS OF JUNCTION DIODES

In this section we study the characteristics of real diodes—specifically, semiconductor junction diodes made of silicon. The physical processes that give rise to the diode terminal characteristics, and to the name “junction diode,” will be studied in the next section.

Figure 3.7 shows the i – v characteristic of a silicon junction diode. The same characteristic is shown in Fig. 3.8 with some scales expanded and others compressed, so as to reveal details. Note that the scale changes have resulted in the apparent discontinuity at the origin.

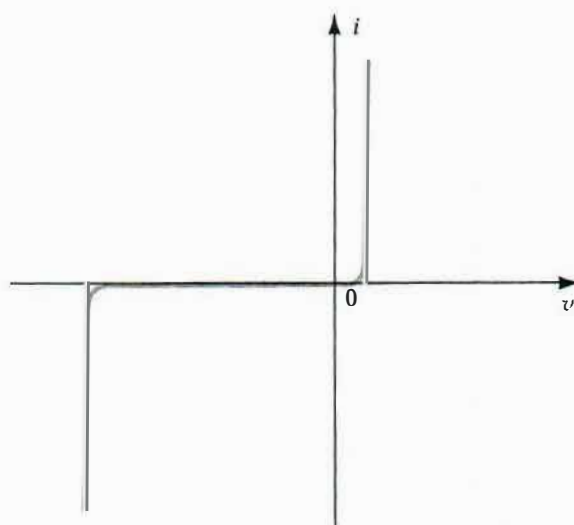


Fig. 3.7 The i – v characteristic of a silicon junction diode.

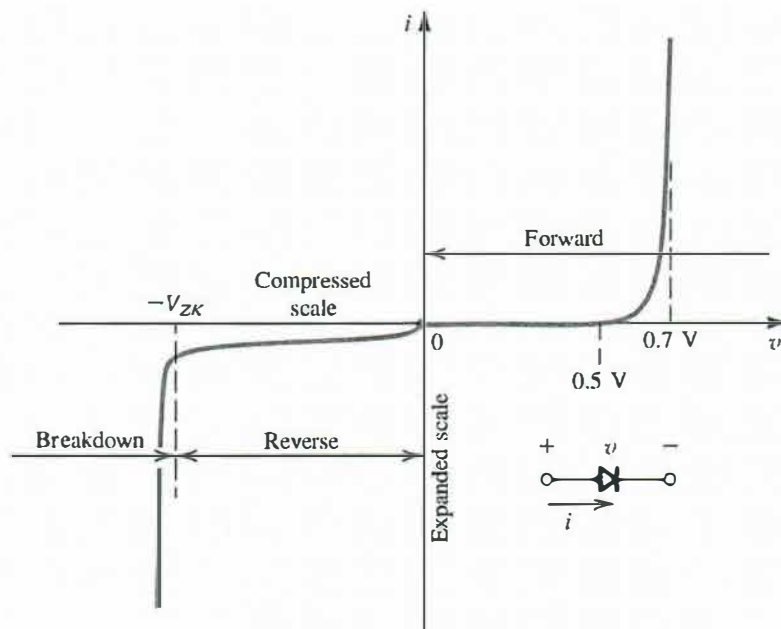


Fig. 3.8 The diode i - v relationship with some scales expanded and others compressed in order to reveal details.

As indicated, the characteristic curve consists of three distinct regions:

1. The forward-bias region, determined by $v > 0$
2. The reverse-bias region, determined by $v < 0$
3. The breakdown region, determined by $v < -V_{ZK}$

These three regions of operation are described in the following.

The Forward-Bias Region

The forward-bias—or, simply, forward—region of operation is entered when the terminal voltage v is positive. In the forward region the i - v relationship is closely approximated by

$$i = I_S(e^{v/nV_T} - 1) \quad (3.1)$$

In this equation I_S is a constant for a given diode at a given temperature. A formula for I_S in terms of the diode's physical parameters and temperature will be given in the next section. The current I_S is usually called the **saturation current** (for reasons that will become apparent shortly). Another name for I_S , one that we will occasionally use, is the **scale current**. This name arises from the fact that I_S is directly proportional to the cross-sectional

area of the diode. Thus doubling of the junction area results in a diode with double the value of I_S and, as the diode equation indicates, double the value of current i for a given forward voltage v . For “small-signal” diodes, which are small-size diodes intended for low-power applications, I_S is of the order of 10^{-15} A. The value of I_S is, however, a very strong function of temperature. As a rule of thumb, I_S doubles in value for every 5°C rise in temperature.¹

The voltage V_T in Eq. (3.1) is a constant called the **thermal voltage**, given by

$$V_T = \frac{kT}{q} \quad (3.2)$$

where

k = Boltzmann's constant = 1.38×10^{-23} joules/kelvin

T = the absolute temperature in kelvins = $273 + \text{temperature in } ^\circ\text{C}$

q = the magnitude of electronic charge = 1.60×10^{-19} coulomb

At room temperature (20°C) the value of V_T is 25.2 mV. In rapid, approximate circuit analysis we shall use $V_T \approx 25$ mV at room temperature.²

In the diode equation the constant n has a value between 1 and 2, depending on the material and the physical structure of the diode. Diodes made using the standard integrated-circuit fabrication process exhibit $n = 1$ when operated under normal conditions.³ Diodes available as discrete two-terminal components generally exhibit $n = 2$. In general, we shall assume $n = 1$ unless otherwise specified.

For appreciable current i in the forward direction, specifically for $i \gg I_S$, Eq. (3.1) can be approximated by the exponential relationship

$$i \approx I_S e^{v/nV_T} \quad (3.3)$$

This relationship can be expressed alternatively in the logarithmic form

$$v = nV_T \ln \frac{i}{I_S} \quad (3.4)$$

where $\ln$ denotes the natural (base e) logarithm.

The exponential relationship of the current i to the voltage v holds over many decades of current (a span of as many as seven decades—that is, a factor of 10^7 —can be found). This is quite a remarkable property of junction diodes, one that is also found in bipolar junction transistors and one that has been exploited in many interesting applications.

¹ An excellent discussion of the temperature dependence of diode characteristics is given in Hodges and Jackson (1988), pp. 146–148. Also given is a derivation for the temperature coefficient of I_S .

² A slightly higher ambient temperature (25°C or so) is usually assumed for electronic equipment operating inside a cabinet. At this temperature, $V_T \approx 25.8$ mV. Nevertheless, for the sake of simplicity and to promote rapid circuit analysis, we shall use the more convenient value of $V_T \approx 25$ mV throughout this book.

³ In an integrated circuit, diodes are usually obtained by connecting a bipolar junction transistor (BJT) as a two-terminal device, as will be seen in Chapter 4.

Let us consider the forward i - v relationship in Eq. (3.3) and evaluate the current I_1 corresponding to a diode voltage V_1 :

$$I_1 = I_S e^{V_1/nV_T}$$

Similarly, if the voltage is V_2 , the diode current I_2 will be

$$I_2 = I_S e^{V_2/nV_T}$$

These two equations can be combined to produce

$$\frac{I_2}{I_1} = e^{(V_2 - V_1)/nV_T}$$

which can be rewritten

$$V_2 - V_1 = nV_T \ln \frac{I_2}{I_1}$$

or, in terms of base-10 logarithms,

$$V_2 - V_1 = 2.3nV_T \log \frac{I_2}{I_1} \quad (3.5)$$

This equation simply states that for a decade (factor of 10) change in current, the diode voltage drop changes by $2.3nV_T$, which is approximately 60 mV for $n = 1$ and 120 mV for $n = 2$. This also suggests that the diode i - v relationship is most conveniently plotted on semilog paper. Using the vertical linear axis for v and the horizontal, log axis for i , one obtains a straight line with a slope of $2.3nV_T$ per decade of current. Finally, it should be mentioned that not knowing the exact value of n (which can be obtained from a simple experiment), circuit designers use the convenient approximate number of 0.1 V/decade for the slope of the diode logarithmic characteristic.

A glance at the i - v characteristic in the forward region (Fig. 3.8) reveals that the current is negligibly small for v smaller than about 0.5 V. This value is usually referred to as the **cut-in voltage**. It should be emphasized, however, that this apparent threshold in the characteristic is simply a consequence of the exponential relationship. Another consequence of this relationship is the rapid increase of i . Thus for a "fully conducting" diode, the voltage drop lies in a narrow range, approximately 0.6 to 0.8 V. This gives rise to a simple "model" for the diode where it is assumed that a conducting diode has approximately a 0.7-V drop across it. Diodes with different current ratings (that is, different areas and correspondingly different I_S) will exhibit the 0.7-V drop at different currents. For instance, a small-signal diode may be considered to have a 0.7-V drop at $i = 1$ mA, while a higher-power diode may have a 0.7-V drop at $i = 1$ A. We will return to the topics of diode-circuit analysis and diode models in Section 3.4.

EXAMPLE 3.3

A silicon diode said to be a 1-mA device displays a forward voltage of 0.7 V at a current of 1 mA. Evaluate the junction scaling constant I_S in the event that n is either 1 or 2. What scaling constants would apply for a 1-A diode of the same manufacture that conducts 1 A at 0.7 V?

SOLUTION

Since

$$i = I_S e^{v/nV_T}$$

then

$$I_S = i e^{-v/nV_T}$$

For the 1-mA diode:

$$\text{If } n = 1: I_S = 10^{-3} e^{-700/25} = 6.9 \times 10^{-16} \text{ A, or about } 10^{-15} \text{ A}$$

$$\text{If } n = 2: I_S = 10^{-3} e^{-700/50} = 8.3 \times 10^{-10} \text{ A, or about } 10^{-9} \text{ A}$$

The diode conducting 1 A at 0.7 V corresponds to 1000 1-mA diodes in parallel with a total junction area 1000 times greater. Thus I_S is also 1000 times greater, being 1 pA and 1 μ A, respectively for $n = 1$ and $n = 2$.

From this example it should be apparent that the value of n used can be quite important.

Since both I_S and V_T are functions of temperature, the forward i - v characteristic varies with temperature as illustrated in Fig. 3.9. At a given constant diode current the voltage drop across the diode decreases by approximately 2 mV for every 1°C increase in temperature. The change in diode voltage with temperature has been exploited in the design of electronic thermometers.

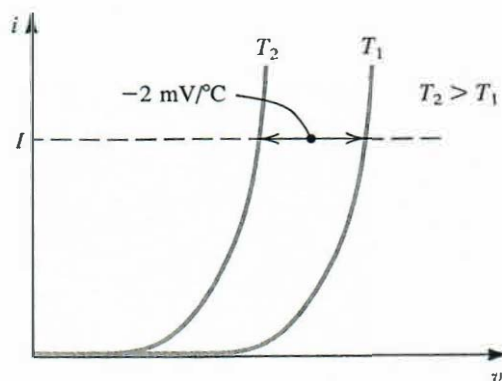


Fig. 3.9 Illustrating the temperature dependence of the diode forward characteristic. At a constant current, the voltage drop decreases by approximately 2 mV for every 1°C increase in temperature.

Exercises

3.6 Consider a silicon diode with $n = 1.5$. Find the change in voltage if the current changes from 0.1 mA to 10 mA.

Ans. 172.5 mV

3.7 A silicon junction diode with $n = 1$ has $v = 0.7$ V at $i = 1$ mA. Find the voltage drop at $i = 0.1$ mA and $i = 10$ mA.

Ans. 0.64 V; 0.76 V

3.8 Using the fact that a silicon diode has $I_S = 10^{-14}$ A at 25°C and that I_S increases by 15% per °C rise in temperature, find the value of I_S at 125°C.

Ans. 1.17×10^{-8} A

The Reverse-Bias Region

The reverse-bias region of operation is entered when the diode voltage v is made negative. Equation (3.1) predicts that if v is negative and a few times larger than V_T (25 mV) in magnitude, the exponential term becomes negligibly small compared to unity and the diode current becomes

$$i \approx -I_S$$

that is, the current in the reverse direction is constant and equal to I_S . This constancy is the reason behind the term *saturation current*.

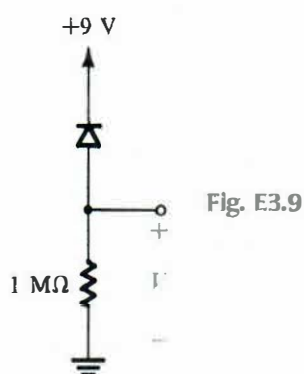
Real diodes exhibit reverse currents that, though quite small, are much larger than I_S . For instance, a small-signal diode whose I_S is of the order of 10^{-14} to 10^{-15} A could show a reverse current of the order of 1 nA. The reverse current also increases somewhat with the increase in magnitude of the reverse voltage. Note that because of the very small magnitude of the current these details are not clearly evident on the diode i - v characteristic of Fig. 3.8.

A good part of the reverse current is due to leakage effects. These leakage currents are proportional to the junction area, just as I_S is. Their dependence on temperature, however, is different from that of I_S . Thus, whereas I_S doubles for every 5°C rise in temperature, the corresponding rule of thumb for the temperature dependence of the reverse current is that it doubles for every 10°C rise in temperature.

$$i = I_s (e^{v/mV_T} - 1) \quad (3.1)$$

Exercise

3.9 The diode in the circuit of Fig. E3.9 is a large, high-current device whose reverse leakage is reasonably independent of voltage. If $V = 1$ V at 20°C , find the value of V at 40°C and at 0°C .



Ans. 4 V; 0.25 V

The Breakdown Region

The third distinct region of diode operation is the breakdown region, which can be easily identified on the diode $i-v$ characteristic in Fig. 3.8. The breakdown region is entered when the magnitude of the reverse voltage exceeds a threshold value specific to the particular diode and called the **breakdown voltage**. This is the voltage at the “knee” of the $i-v$ curve in Fig. 3.8 and is denoted V_{ZK} , where the subscript Z stands for zener (to be explained shortly) and K denotes knee.

As can be seen from Fig. 3.8, in the breakdown region the reverse current increases rapidly, with the associated increase in voltage drop being very small. Diode breakdown is normally not destructive provided that the power dissipated in the diode is limited by external circuitry to a “safe” level. This safe value is normally specified on the device data sheets. It therefore is necessary to limit the reverse current in the breakdown region to a value consistent with the permissible power dissipation.

The fact that the diode $i-v$ characteristic in breakdown is almost a vertical line enables it to be used in voltage regulation. This subject will be studied in Section 3.6.

3.3 PHYSICAL OPERATION OF DIODES

Having studied the terminal characteristics of junction diodes, we will now briefly consider the physical processes that give rise to these characteristics. The following treatment of device physics is somewhat simplified; nevertheless, it should provide sufficient background for the design of diode and other semiconductor circuits.

CHAPTER 13

MOS Digital Circuits

13.

	Introduction	13.9 Semiconductor Memories: Types and Architectures
13.1	Digital Circuit Design: An Overview	13.10 Random-Access Memory (RAM) Cells
13.2	Design and Performance Analysis of the CMOS Inverter	13.11 Sense Amplifiers and Address Decoders
13.3	CMOS Logic-Gate Circuits	13.12 Read-Only Memory (ROM)
13.4	Pseudo-NMOS Logic Circuits	13.13 SPICE Simulation Example
13.5	Pass-Transistor Logic Circuits	Summary
13.6	Dynamic Logic Circuits	Bibliography
13.7	Latches and Flip-Flops	Problems
13.8	Multivibrator Circuits	

INTRODUCTION

This chapter is concerned with the study of MOS digital integrated circuits, by far the most popular technology for the implementation of digital systems. The small size, ease of fabrication, and low power dissipation of MOSFETs enable extremely high levels of integration of both logic and memory circuits.

The chapter begins with an overview section with the objective of placing in proper perspective the material we shall study in this chapter and the next. Then, building on the study of the CMOS inverter in Section 5.8, we take a comprehensive look at its design and analysis. This material is then applied to the design of CMOS logic circuits and two other types of logic circuits (namely, pseudo-NMOS and pass-transistor logic) that are frequently employed in special applications as a supplement to CMOS.

To reduce the power dissipation even further, and simultaneously to increase performance (speed of operation), dynamic logic techniques are employed. This challenging topic is the subject of Section 13.6 and completes our study of combinational circuits. The remainder of the chapter is concerned with the equally important subject of sequential circuits. These include the basic bistable latch and a variety of flip-flops based on it, pulse generators

and pulse-shaping circuits, and, most significantly, memory circuits (both RAMs and ROMs). The chapter concludes with a SPICE simulation example.

In summary, this chapter provides a reasonably comprehensive and in-depth treatment of MOS digital integrated circuit design, perhaps the most significant area (at least in terms of production volume and societal impact) of electronic circuits. To gain the most out of studying this chapter, the reader must be thoroughly familiar with the MOS transistor. Thus, a review of Chapter 5 is recommended, and a careful study of Section 5.8 is a must!

13.1 DIGITAL CIRCUIT DESIGN: AN OVERVIEW

In this section, we build on the introduction to digital circuits presented in Section 1.7 and provide an overview of the subject. We discuss the various technologies and logic-circuit families currently in use, consider the parameters employed to characterize the operation and performance of logic circuits, and finally mention the various styles for digital-system design.

13.1.1 Digital IC Technologies and Logic-Circuit Families

The chart in Figure 13.1 shows the major IC technologies and logic-circuit families that are currently (1997) in use. The concept of a logic-circuit family perhaps needs a few words of explanation. Members of each family are made with the same technology, have similar circuit structure, and exhibit the same basic features. Each logic-circuit family offers a unique set of advantages and disadvantages. In the conventional style of designing systems, one selects an appropriate logic family (e.g., TTL, CMOS, or ECL) and attempts to implement as much of the system as possible using circuit modules (packages) that belong to this family. In this way, interconnection of the various packages is relatively straightforward. If, on the other hand, packages from more than one family are used, one has to design suitable *interface circuits*. The selection of a logic family is based on such considerations as logic flexibility, speed of operation, availability of complex functions, noise immunity, operating-temperature range, power dissipation, and cost. We will discuss some of these considerations in this chapter and the next. To begin with, we make some brief remarks on each of the four technologies listed in the chart of Fig. 13.1.

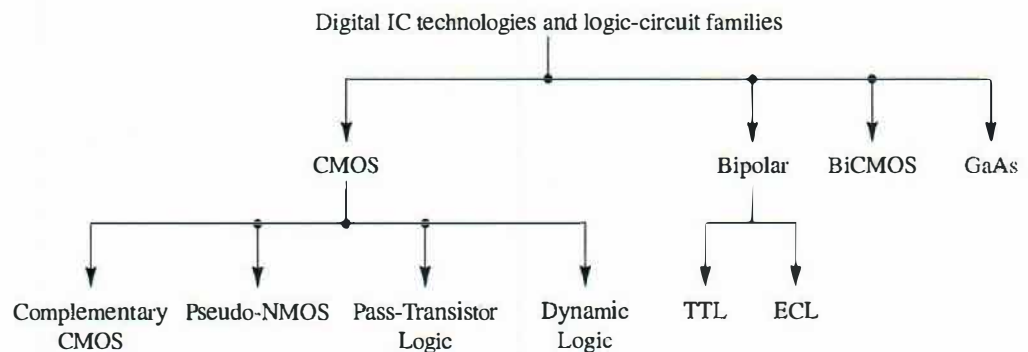


Fig. 13.1 Digital IC technologies and logic-circuit families.

CMOS. Although shown as one of four possible technologies, this is not an indication of its share of the digital IC market; CMOS technology is by a large margin the most dominant of all the IC technologies available for digital-circuit design. As mentioned earlier, CMOS has replaced NMOS, which was employed in the early days of VLSI in the 1970s. There are a number of reasons for this development, the most important of which is the much lower power dissipation of CMOS circuits. CMOS has also replaced bipolar as the technology of choice in digital-system design and has made possible levels of integration (or circuit packing densities) and a range of applications that would never have been possible with bipolar technology. Furthermore, CMOS continues to advance, whereas there appear to be few innovations at the present time in bipolar digital circuits.¹ Some of the reasons for CMOS displacing bipolar technology in digital applications are:

1. CMOS logic circuits dissipate much less power than bipolar logic circuits, and thus one can pack more CMOS circuits on a chip than is possible with bipolar circuits. We will have a lot more to say about power dissipation in the following sections.
2. The high input impedance of the MOS transistor allows the designer to use charge storage as a means for the temporary storage of information in both logic and memory circuits. This technique cannot be used in bipolar circuits.
3. The feature size (that is, minimum channel length) of the MOS transistor has decreased dramatically over the years, with some recently reported designs utilizing channel lengths as short as $0.15\text{ }\mu\text{m}$. This permits very tight circuit packing and, correspondingly, very high levels of integration.

Of the various forms of CMOS, complementary CMOS circuits based on the inverter studied in Section 5.8 are the most widely used. They are available both as small-scale integrated (SSI) circuit packages (containing 1 to 10 logic gates) and medium-scale integrated (MSI) circuit packages (10 to 100 gates per chip) for assembling digital systems on printed-circuit boards. More significant, complementary CMOS is used in VLSI logic (millions of gates per chip) and memory circuit design. In some applications, complementary CMOS is supplemented by one of two other MOS logic circuit forms. These are pseudo-NMOS, so named because of the similarity of its structure to NMOS, and pass-transistor logic, both of which will be studied in this chapter.

A fourth type of CMOS logic circuit utilizes dynamic techniques to obtain faster circuit operation while keeping the power dissipation very low. Dynamic CMOS logic represents an area of growing importance. Lastly, CMOS technology is used in the design of memory chips, as will be explained in later sections of this chapter.

Bipolar. Two logic-circuit families based on the bipolar junction transistor are in use at present: TTL and ECL. Transistor-transistor logic (TTL or T^2L) was for many years the most widely used logic-circuit family. Its decline was precipitated by the advent of the VLSI

¹ In support of this statement, we offer the following anecdotal evidence: Every year, the November issue of the *IEEE Journal of Solid-State Circuits* is devoted to papers on digital logic and memory circuits. The latest issue available, November 1996, includes 28 papers of which 26 use CMOS technology and 2 utilize BiCMOS (one for a memory chip and the other for a control circuit of disk drives). None of the papers is on bipolar digital circuits!

era. TTL manufacturers, however, fought back with the introduction of low-power and high-speed versions. In these newer versions, the higher speeds of operation are made possible by preventing the BJT from saturating and thus avoiding the slow turn-off process of a saturated transistor. This subject will be studied in Chapter 14. These newer versions utilize the Schottky diode discussed in Chapter 3 and are called Schottky TTL or variations of this name. Because TTL remains a viable digital-system design technology, especially for digital systems that are assembled from SSI and MSI packages, and because of its historical dominance and influence on the entire area of digital IC design, we shall study TTL in some detail in Chapter 14.

The other bipolar logic-circuit family in present use is emitter-coupled logic (ECL). It is based on the current-switch implementation of the inverter, discussed in Section 1.7. The basic element of ECL is the differential pair studied in Chapter 6. Because in ECL, transistors are not allowed to saturate, and it is basically a current-steering circuit (hence the alternative name, current-mode logic), very high speeds of operation are possible. Indeed, of all the commercially available logic-circuit families, ECL is the fastest. ECL is also used in VLSI circuit design when very high operating speeds are required and the designer is willing to accept higher power dissipation and increased silicon area. As such, ECL is considered an important speciality technology.

BiCMOS. BiCMOS combines the high operating speeds possible with BJTs (because of their inherently higher transconductance) with the low power dissipation and other excellent characteristics of CMOS. Like CMOS, BiCMOS allows for the implementation of both analog and digital circuits on the same chip. (See the discussion of analog BiCMOS circuits in Chapters 6 and 10.) At present, BiCMOS is used to great advantage in special applications, including memory chips, where its high performance justifies the more complex process technology it requires.

Gallium Arsenide (GaAs). Gallium arsenide (GaAs) devices were introduced in Section 5.12. As mentioned there, the high carrier mobility in GaAs results in very high speeds of operation. This has been demonstrated in a number of digital IC chips utilizing GaAs technology. The design of logic circuits utilizing GaAs technology will be studied in Chapter 14. It should be pointed out, however, that GaAs remains an "emerging technology," one that appears to have great potential, but has not yet achieved such potential commercially.

13.1.2 Logic-Circuit Characterization

The following parameters are usually used to characterize the operation and performance of a logic-circuit family.

Noise Margins. The static operation of a logic-circuit family is characterized by the voltage-transfer characteristic (VTC) of its basic inverter. Figure 13.2 shows such a VTC and defines its four parameters; V_{OH} , V_{OL} , V_{IH} , and V_{IL} . Note that V_{IH} and V_{IL} are defined as the points at which the slope of the VTC is -1 . Also indicated is the definition of the threshold voltage V_M , or V_{th} as we shall frequently call it, as the point at which $v_O = v_I$. Recall that we discussed the VTC in its generic form in Section 1.7, and have seen actual VTCs in Section 4.14 for the BJT inverter and in Section 5.8 for the CMOS inverter.

The **robustness** of a logic-circuit family is determined by its ability to reject noise and

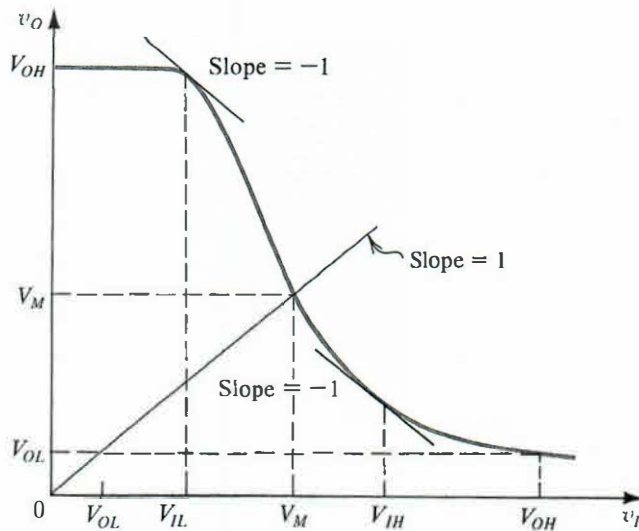


Fig. 13.2 Typical voltage transfer characteristic (VTC) of a logic inverter, illustrating the definition of the critical points.

thus by the noise margins NH_H and NM_L ,

$$NM_H \equiv V_{OH} - V_{IH} \quad (13.1)$$

$$NM_L \equiv V_{IL} - V_{OL} \quad (13.2)$$

An ideal inverter is one for which $NM_H = NM_L = V_{DD}/2$, where V_{DD} is the power-supply voltage. Further, for an ideal inverter, the threshold voltage $V_M = V_{DD}/2$.

Propagation Delay. The dynamic performance of a logic-circuit family is characterized by the propagation delay of its basic inverter. Figure 13.3 illustrates the definition of the low-to-high propagation delay (t_{PLH}) and the high-to-low propagation delay (t_{PHL}). The inverter propagation delay t_P is defined as the average of these two quantities:

$$t_P \equiv \frac{1}{2}(t_{PLH} + t_{PHL}) \quad (13.3)$$

Obviously, the shorter the propagation delay, the higher the speed at which the logic-circuit family can be operated.

Power Dissipation. Power dissipation is an important issue in digital circuit design. The need to minimize the gate power dissipation is motivated by the desire to pack an ever-increasing number of gates on a chip, which in turn is motivated by space and economic considerations. In general, however, modern digital systems utilize large numbers of gates and memory cells, and thus to keep the total power requirement within reasonable bounds, the power dissipation per gate and per memory cell should be kept as low as possible. This is particularly the case for portable, battery-operated equipment such as cellular phones and personal digital assistants (PDAs).

There are two types of power dissipation in a logic gate: static and dynamic. Static power refers to the power the gate dissipates in the absence of switching action. It results from the presence of a path in the gate circuit between the power supply and ground in one or both of its two states (i.e., with the output either low or high). Dynamic power, on the

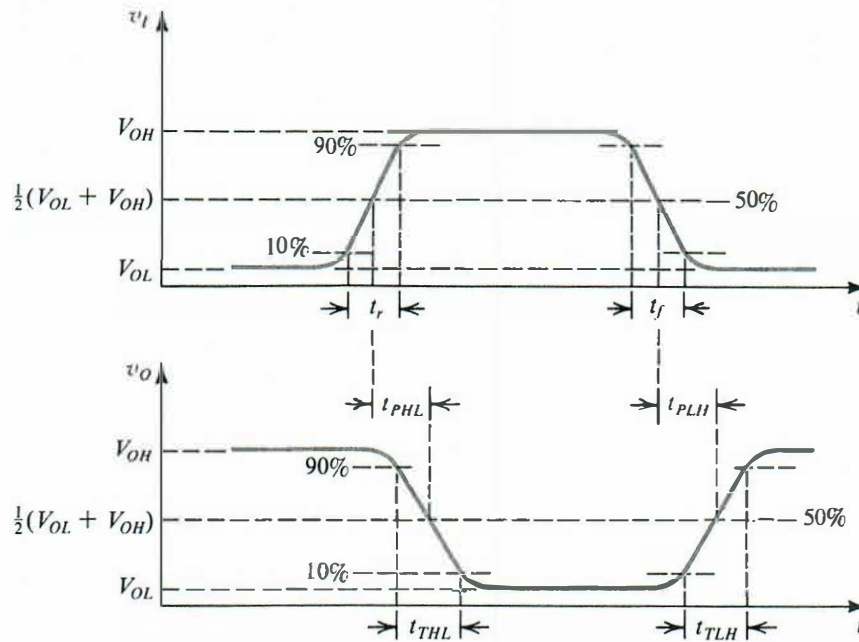


Fig. 13.3 Definitions of propagation delays and switching times of the logic inverter.

other hand, occurs only when the gate is switched: An inverter operated from a power supply V_{DD} , and driving a load capacitance C , dissipates dynamic power P_D ,

$$P_D = fCV_{DD}^2 \quad (13.4)$$

where f is the frequency at which the inverter is being switched. The derivation of this formula (Section 5.8) is based on the assumption that the low and high output voltage levels are 0 and V_{DD} , respectively.

Delay-Power Product. One is usually interested in high-speed performance (low t_P) combined with low power dissipation. Unfortunately, these two requirements are often in conflict; generally, when designing a gate, if one attempts to reduce power dissipation by decreasing the supply voltage, the supply current, or both, the current-driving capability of the gate decreases. This in turn results in longer times to charge and discharge the load and parasitic capacitances, and thus the propagation delay increases. It follows that a figure-of-merit for comparing logic-circuit technologies or families is the delay-power product, defined as

$$DP = P_D t_P \quad (13.5)$$

where P_D is the power dissipation of the gate. Note that DP has the units of joules. The lower the DP figure for a family, the more effective the logic family is.

Silicon Area. An obvious objective in the design of digital VLSI circuits is the minimization of silicon area per logic gate. Smaller area requirement enables the fabrication of a larger number of gates per chip, which has economic and space advantages from a system-

design standpoint. Area reduction occurs in three different ways: through advances in processing technology that enable the reduction of the minimum device size, through advances in circuit design techniques, and through careful chip layout. In this book, our interest lies in circuit design, and we shall make frequent comments on the relationship between the circuit design and its silicon area. As a general rule, the simpler the circuit, the smaller the area it requires. As will be seen shortly, the circuit designer has to decide on device sizes. Choosing smaller devices has the obvious advantage of requiring smaller silicon area and at the same time reducing parasitic capacitances and thus increasing speed. Smaller devices, however, have lower current driving capability, which tends to increase delay. Thus, as in all engineering design problems, there is a trade-off that has to be quantified and exercised in a manner that optimizes whatever aspect of the design is thought to be critical for the application at hand.

Fan-In and Fan-Out. The fan-in of a gate is the number of its inputs. Thus, a four-input NOR gate has a fan-in of 4. Fan-out is the maximum number of similar gates that a gate can drive while remaining within guaranteed specifications. As an example, we saw in Section 4.14 that increasing the fan-out of the BJT inverter reduces V_{OH} and hence NM_H . To keep NM_H above a certain minimum, the fan-out has to be limited to a calculable maximum value.

13.1.3 Styles for Digital System Design

The conventional approach to designing digital systems consists of assembling the system using standard IC packages of various levels of complexity (and hence integration). Many systems have been built this way using, for example, TTL SSI and MSI packages. The advent of VLSI, in addition to providing the system designer with more powerful off-the-shelf components such as microprocessors and memory chips, has made possible alternative design styles. One such alternative is to opt for implementing part or all of the system using one or more *custom VLSI* chips. However, custom IC design is usually economically justified only when the production volume is large (greater than about 100,000 parts).

An intermediate approach, known as *semicustom design*, utilizes *gate-array* chips. These are integrated circuits containing 100,000 or more unconnected logic gates. Their interconnection can be achieved by a final metallization step (performed at the IC fabrication facility) according to a pattern specified by the user so as to implement the user's particular function. A more recently available type of gate array, known as a *field-programmable gate array* (FPGA), can, as its name indicates, be programmed directly by the user. FPGAs provide a very convenient means for the digital-system designer to implement complex logic functions in VLSI form without having to incur either the cost or the "turnaround time" inherent in custom and, to a lesser extent, in semicustom IC design [see Brown and Rose (1996)].

13.1.4 Design Abstraction and Computer Aids

The design of very complex digital systems whether on a single IC or using off-the-shelf components is made possible by the use of different levels of design abstraction and the use of a variety of computer aids. To appreciate the concept of design abstraction consider

the process of designing a digital system using off-the-shelf packages of logic gates. The designer consults data sheets and books to determine the input and output characteristics of the gates, their fan-in and fan-out limitations, and so on. In connecting the gates, the designer needs to adhere to a set of rules specified by the manufacturer in the data sheets. The designer does not need to consider, in a direct way, the circuit inside the gate package. In effect, the circuit has been abstracted in the form of a functional block that can be used as a component. This greatly simplifies system design. The digital-IC designer follows a similar process. Circuit blocks are designed, characterized, and stored in a library as *standard cells*. These cells can then be used by the IC designer to assemble a larger subsystem (e.g., an adder or a multiplier), which in turn is characterized and stored as a functional block to be used in the design of an even larger system.

At every level of design abstraction, the need arises for simulation and other computer programs that help make the design process as automated as possible. Whereas SPICE is employed in circuit simulation, other software tools are utilized at other levels and in other phases of the design process. Although digital-system design and design automation are outside the scope of this book, it is important that the reader appreciate the role of design abstraction and computer aids in digital design. They are what make it humanly possible to design a 10-million-transistor digital IC. Unfortunately, analog IC design does not lend itself to the same level of abstraction and automation. Each analog IC to a large extent has to be “handcrafted.” As a result, the complexity and density of analog ICs remain much less than that possible within a digital IC.

Whatever approach or style is adopted in digital design, some familiarity with the various digital-circuit technologies and design techniques is essential. This chapter and the next aim to provide such a background.

13.2 DESIGN AND PERFORMANCE ANALYSIS OF THE CMOS INVERTER

The CMOS logic inverter was introduced and studied in Section 5.8, which we urge the reader to review before proceeding any further. In this section, we take a more comprehensive look at the inverter, investigating its performance and exploring the trade-offs available in its design. This material will serve as the foundation for the study of CMOS logic circuits in the following section.

Circuit Structure

The inverter circuit, shown in Fig. 13.4(a), consists of a pair of complementary MOSFETs switched by the input voltage v_I . Although not shown, the source of each device is connected to its body, thus eliminating the body effect. Usually, the threshold voltages V_{tn} and V_{tp} are equal in magnitude; that is, $V_{tn} = |V_{tp}| = V_t$, which is in the range of 0.2 to 1 V, with values near the lower end of this range for modern process technologies with small feature size (e.g., with channel length of 0.3–0.5 μm).

The inverter circuit can be represented by a pair of switches operated in a complementary fashion, as shown in Fig. 13.4(b). As indicated, each switch is modeled by a finite on resistance, which is the source–drain resistance of the respective transistor, evaluated near

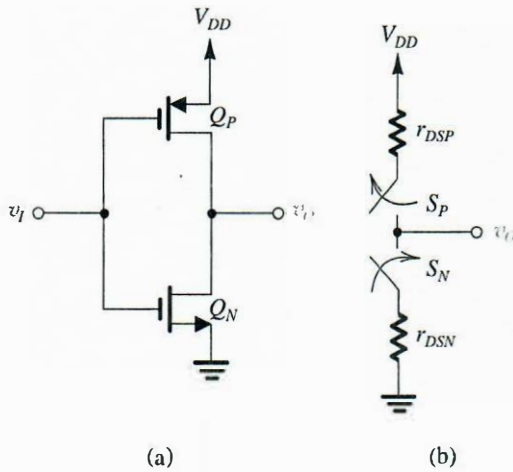


Fig. 13.4 (a) The CMOS inverter and (b) its representation as a pair of switches operated in a complementary fashion.

$$|v_{DS}| = 0,$$

$$r_{DSN} = 1 / \left[k'_n \left(\frac{W}{L} \right)_n (V_{DD} - V_t) \right] \quad (13.6)$$

$$r_{DSP} = 1 / \left[k'_p \left(\frac{W}{L} \right)_p (V_{DD} - V_t) \right] \quad (13.7)$$

Static Operation

With $v_I = 0$, $v_O = V_{OH} = V_{DD}$, and the output node is connected to V_{DD} through the resistance r_{DSP} of the pull-up transistor Q_P . Similarly, with $v_I = V_{DD}$, $v_O = V_{OL} = 0$, and the output node is connected to ground through the resistance r_{DSN} of the pull-down transistor Q_N . Thus, in steady state, no direct-current path exists between V_{DD} and ground, and the static current and the static power dissipation are both zero (leakage effects are usually negligibly small).

The voltage transfer characteristic (VTC) of the inverter is shown in Fig.13.5, from which it is confirmed that the output voltage levels are 0 and V_{DD} , and thus the output voltage swing is the maximum possible. The fact that V_{OL} and V_{OH} are independent of device dimensions makes CMOS very different from other forms of MOS logic (e.g., NMOS circuits utilizing enhancement or depletion load devices, which are based on the amplifier circuits studied in Sections 5.7.4). For these NMOS circuits, V_{OL} depends on the ratio of the transconductance parameters of the devices, that is, on the ratio $\left(k' \frac{W}{L} \right)_{\text{driver}} / \left(k' \frac{W}{L} \right)_{\text{load}}$. Such circuits are therefore known as *ratioed* logic circuits; correspondingly, CMOS is said to belong to the *ratioless* type.

The CMOS inverter can be made to switch at the midpoint of the logic swing, 0 to V_{DD} , that is, at $V_{DD}/2$, by appropriately *sizing the transistors*. Specifically, it can be shown

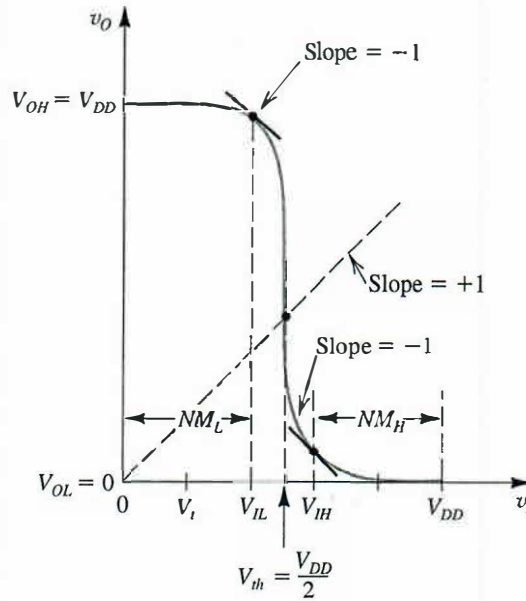


Fig. 13.5 The voltage transfer characteristic (VTC) of the CMOS inverter when Q_N and Q_P are matched.

that the switching threshold is given by

$$V_{th} = \frac{V_{DD} - |V_{tp}| + \sqrt{k_n/k_p} V_{tn}}{1 + \sqrt{k_n/k_p}} \quad (13.8)$$

where $k_n = k'_n(W/L)_n$ and $k_p = k'_p(W/L)_p$, from which we see that for the typical case where $V_{tn} = |V_{tp}|$, $V_{th} = V_{DD}/2$ for $k_n = k_p$, that is

$$k'_n(W/L)_n = k'_p(W/L)_p \quad (13.9)$$

Thus a symmetrical transfer characteristic is obtained when the devices are designed to have equal transconductance parameters, a condition we refer to as *matching*. Since μ_n is two to three times larger than μ_p , matching is achieved by making $(W/L)_p$ two to three times (i.e., μ_n/μ_p times) $(W/L)_n$,

$$\left(\frac{W}{L}\right)_p = \frac{\mu_n}{\mu_p} \left(\frac{W}{L}\right)_n \quad (13.10)$$

Normally, the two devices have the same channel length, L , which is set at the minimum allowable for the given process technology. The minimum width of the NMOS transistor is usually one and a half to two times L , and the width of the PMOS transistor two to three times that. For example, for a 1.2- μm process for which $\mu_n/\mu_p = 3$, $L = 1.2 \mu\text{m}$, $(W/L)_n = 1.8/1.2$, and $(W/L)_p = 5.4/1.2$. As we shall discuss shortly, if the inverter is required to drive a relatively large capacitive load, the transistors are made wider. However, to conserve chip area, most of the inverters would have this “minimum size.” For future purposes, we shall denote the (W/L) ratio of the NMOS transistor of this minimum-size inverter by n and the (W/L) ratio of the PMOS transistor by p . Since the inverter area can be represented by $W_n L_n + W_p L_p = (W_n + W_p)L$, the area of the minimum-size inverter

is $(n + p)L^2$, and we can use the factor $(n + p)$ as proxy for area. For the example cited earlier, $n = 1.5$, $p = 4.5$, and the area factor $n + p = 6$.

Besides placing the gate threshold at the center of the logic swing, matching the transconductance parameters of Q_N and Q_P provides the inverter with equal current-driving capability in both directions (pull-up and pull-down). Furthermore, and obviously related, it makes $r_{DSN} = r_{DSP}$. Thus an inverter with matched transistors will have equal propagation delays, t_{PLH} and t_{PHL} .

When the inverter threshold is at $V_{DD}/2$, the noise margins NM_H and NM_L are equalized and their values are maximized, such that:

$$NM_H = NM_L = \frac{3}{8} \left(V_{DD} + \frac{2}{3} V_t \right) \quad (13.11)$$

Since typically $V_t = 0.1$ to $0.2 V_{DD}$, the noise margins are approximately $0.4 V_{DD}$. This value, being close to half the power supply voltage, makes the CMOS inverter nearly ideal from a noise-immunity standpoint. Further, since the inverter dc input current is practically zero, the noise margins are not dependent on the gate fan-out.

Although we have emphasized the advantages of matching Q_N and Q_P , there are occasions in which this scaling is not adopted. One might, for instance, forego the advantages of matching in return for reducing chip area, and simply make $(W/L)_p = (W/L)_n$. There are also instances in which a deliberate mismatch is used so as to place V_{th} at a specified value other than $V_{DD}/2$. Note that by making $k_n > k_p$, V_{th} moves closer to zero, whereas $k_p > k_n$ moves V_{th} closer to V_{DD} .

As a final comment on the inverter VTC, we note that the slope in the transition region, though large, is finite and is given by $-(g_{mN} + g_{mP})(r_{oN}/r_{oP})$.

Dynamic Operation

The propagation delay of the inverter is usually determined under the condition that it is driving an identical inverter. This situation is depicted in Fig. 13.6. We wish to analyze this circuit to determine the propagation delay of the inverter comprising Q_1 and Q_2 , which is driven by a low-impedance source v_I , and is loaded by the inverter comprising Q_3 and Q_4 . Indicated in the figure are the various transistor internal capacitances that are connected to the output node of the (Q_1, Q_2) inverter. Obviously, an exact pencil-and-paper analysis of this circuit will be too complicated to yield useful design insight, and a simplification of the circuit is in order. Specifically, we wish to replace all the capacitances attached to the inverter output node with a single capacitance C connected between the output node and ground. If we are able to do that, we can utilize the results of the transient analysis performed in Section 5.8. Toward that end, we note that during t_{PLH} or t_{PHL} , the output of the first inverter changes from 0 to $V_{DD}/2$ or from $V_{DD}/2$ to 0, respectively. It follows that the second inverter remains in the same state during each of our analysis intervals. This observation will have an important bearing on our estimation of the equivalent input capacitance of the second inverter. Let's now consider the contribution of each of the capacitances in Fig. 13.6 to the value of the equivalent load capacitance C :

1. The gate-drain overlap capacitance of Q_1 , C_{gd1} , can be replaced by an equivalent capacitance between the output node and ground of $2C_{gd1}$. The factor 2 arises because of the Miller effect (Section 7.4). Specifically, note that as v_I goes high

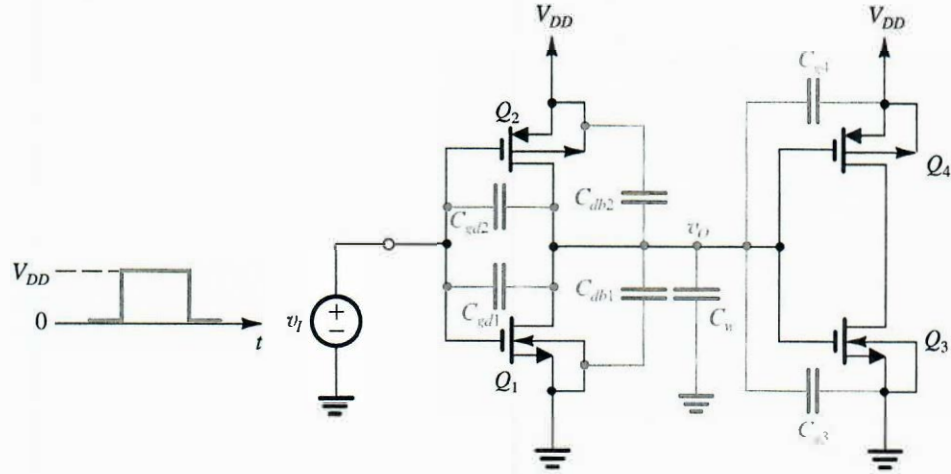


Fig. 13.6 Circuit for analyzing the propagation delay of the inverter. Here, the inverter under study is formed by Q_1 and Q_2 . It is driving an identical inverter formed by Q_3 and Q_4 .

and v_O goes low by the same amount, the change in voltage across C_{gd1} is twice that amount. Thus the output node sees in effect twice the value of C_{gd1} . The same applies for the gate-drain overlap capacitance of Q_2 , C_{gd2} , which can be replaced by a capacitance $2C_{gd2}$ between the output node and ground.

2. Each of the drain-body capacitances C_{db1} and C_{db2} has a terminal at a constant voltage. Thus for the purpose of our analysis here, C_{db1} and C_{db2} can be replaced with equal capacitances between the output node and ground. Note, however, that the formulas given in Section 5.10 for calculating C_{db1} and C_{db2} are small-signal relationships, whereas the analysis here is obviously a large-signal one. A technique has been developed for finding equivalent large-signal values for C_{db1} and C_{db2} [see Hodges and Jackson (1988), and Rabaey (1996)].
3. Since the second inverter does not switch states, we will assume that the input capacitances of Q_3 and Q_4 remain approximately constant and equal to the total gate capacitance WLC_{ox} . That is, the input capacitance of the load inverter will be

$$C_{g3} + C_{g4} = (WL)_3 C_{ox} + (WL)_4 C_{ox}$$

4. The last component of C is the wiring capacitance C_w , which simply adds to the value of C .

Thus, the total value of C is given by

$$C = 2C_{gd1} + 2C_{gd2} + C_{db1} + C_{db2} + C_{g3} + C_{g4} + C_w \quad (13.12)$$

Having determined an approximate value for the equivalent capacitance between the inverter output node and ground, we can utilize the circuits in Figs. 13.7(a) and 13.7(b) to determine t_{PHL} and t_{PLH} , respectively. Since the two circuits are similar, we need only consider one and apply the result directly to the other. Consider the circuit in Fig. 13.7(a), which applies when v_I goes high and Q_N discharges C from its initial voltage of V_{DD} to the final value

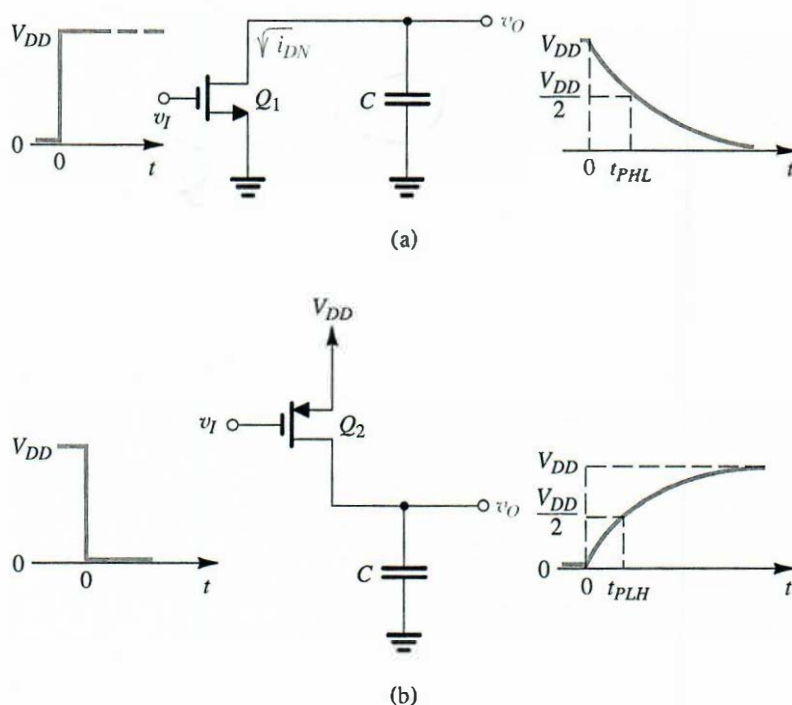


Fig. 13.7 Equivalent circuits for determining the propagation delays (a) t_{PHL} and (b) t_{PLH} of the inverter.

of 0. The analysis is somewhat complicated by the fact that initially Q_N will be in the saturation mode and then, when v_O falls below $V_{DD} - V_t$, it will go into the triode region of operation. We have in fact performed this analysis in Section 5.8 and obtained the following approximate expression for t_{PHL} :

$$t_{PHL} = \frac{1.6C}{k'_n \left(\frac{W}{L}\right)_n V_{DD}} \quad (13.13)$$

where we have assumed that $V_t \cong 0.2 V_{DD}$, which is typically the case.

There is an alternative, an approximate but simpler, method for analyzing the circuit in Fig. 13.7(a). It is based on computing an average value for the discharge current i_{DN} during the interval $t = 0$ to $t = t_{PHL}$. Specifically, at $t = 0$, Q_N will be saturated, and $i_{DN}(0)$ is given by

$$i_{DN}(0) = \frac{1}{2} k'_n \left(\frac{W}{L}\right)_n (V_{DD} - V_t)^2 \quad (13.14)$$

At $t = t_{PHL}$, Q_N will be in the triode region, and $i_{DN}(t_{PHL})$ will be

$$i_{DN}(t_{PHL}) = k'_n \left(\frac{W}{L}\right)_n \left[(V_{DD} - V_t) \frac{V_{DD}}{2} - \frac{1}{2} \left(\frac{V_{DD}}{2}\right)^2 \right] \quad (13.15)$$

The average discharge current can then be found as

$$i_{DN}|_{av} = \frac{1}{2}[i_{DN}(0) + i_{DN}(t_{PHL})] \quad (13.16)$$

and the discharge interval t_{PHL} computed from

$$t_{PHL} = \frac{C \Delta V}{i_{DN}|_{av}} = \frac{C V_{DD}/2}{i_{DN}|_{av}} \quad (13.17)$$

Utilizing Eqs. (13.14) through (13.17) and substituting $V_i \approx 0.2 V_{DD}$ gives

$$t_{PHL} \approx \frac{1.7 C}{k'_n \left(\frac{W}{L}\right)_n V_{DD}} \quad (13.18)$$

which yields a value very close to that obtained by the more precise formula of Eq. (13.13). Which formula to use is not very relevant, for we have already made many approximations. Indeed, our interest in these formulas is not in obtaining a precise value of t_{PHL} , but in what they tell us about the effect of the various elements on determining the inverter delay. It is such insight that the circuit designer hopes to glean from manual analysis. Precise values for delay can be determined using computer simulation (Section 13.13).

An expression for the low-to-high inverter delay, t_{PLH} , can be written by analogy to the t_{PHL} expression in Eq. (13.17),

$$t_{PLH} \approx \frac{1.7 C}{k'_p \left(\frac{W}{L}\right)_p V_{DD}} \quad (13.19)$$

Finally, the propagation delay t_P can be found as the average of t_{PHL} and t_{PLH} ,

$$t_P = \frac{1}{2}(t_{PHL} + t_{PLH})$$

Examination of the formulas in Eqs. (13.18) and (13.19) enables us to make a number of useful observations:

1. As expected, the two components of t_P can be equalized by selecting the (W/L) ratios so as to equalize k_n and k_p , that is, by matching Q_N and Q_P .
2. Since t_P is proportional to C , the designer should strive to reduce C . This is achieved by using the minimum possible channel length and by minimizing wiring and other parasitic capacitances. Careful layout of the chip can result in significant reduction in such capacitances and in the value of C_{db} .
3. Using a process technology with larger transconductance parameter k' can result in shorter propagation delays. Keep in mind, however, that for such processes C_{ox} is increased, and thus the value of C increases at the same time.
4. Using larger (W/L) ratios can result in a reduction in t_P . Care, however, should be exercised here since increasing the size of the devices increases the value of C , and

thus the expected reduction in t_P might not materialize. Reducing t_P by increasing (W/L) , however, is an effective strategy in situations where C is dominated by components not directly related to the size of the driving device.

5. A larger supply voltage V_{DD} results in a lower t_P . However, V_{DD} is determined by the process technology and thus is often not under the control of the designer. Furthermore, modern process technologies in which device sizes are reduced require lower V_{DD} . At the time of this writing (1997), a new standard value for V_{DD} has emerged, namely, 3.3 V (or 3.0 V) replacing the 5-V standard of the past many years. A motivating factor for lowering V_{DD} is the need to keep the dynamic power dissipation at acceptable levels, especially in very-high-density chips. We will have more to say on this point shortly.

These observations clearly illustrate the conflicting requirements and the trade-offs available in the design of a CMOS digital integrated circuit (and indeed in any engineering design problem).

Dynamic Power Dissipation

The negligible static power dissipation of CMOS has been a significant factor in motivating its replacing NMOS as the technology of choice in implementing high-density VLSI circuits. However, as the number of gates per chip steadily increases, the dynamic power dissipation has become a serious issue. The dynamic power dissipated in the CMOS inverter is given by Eq. (13.4), which we repeat here as

$$P_D = f C V_{DD}^2 \quad (13.20)$$

where f is the frequency at which the gate is switched. It follows that minimizing C is an effective means for reducing dynamic power dissipation. An even more effective strategy is the use of a lower power supply voltage. As we mentioned, new CMOS process technologies utilize V_{DD} of 3.3 V. In fact, some recently reported designs employ V_{DD} of 2–2.5 V internal to the chip while keeping the external voltage at the 3.3-V standard value. These newer chips, however, pack much more circuitry on the chip (as many as 10 million transistors) and operate at higher frequencies (microprocessor clock frequencies as high as 500 MHz have recently been reported). The dynamic power dissipation of such high-density chips can be in the range of tens of watts.

EXAMPLE 13.1

Consider a CMOS inverter fabricated in a process for which $C_{ox} = 0.9 \text{ fF}/\mu\text{m}^2$, $\mu_n C_{ox} = 50 \mu\text{A}/\text{V}^2$, $\mu_p C_{ox} = 20 \mu\text{A}/\text{V}^2$, $V_{tn} = -V_{tp} = 1 \text{ V}$, and $V_{DD} = 5 \text{ V}$. The W/L ratio of Q_N is $4 \mu\text{m}/2 \mu\text{m}$, and that for Q_P is $10 \mu\text{m}/2 \mu\text{m}$. The gate-drain overlap capacitance is specified to be $0.5 \text{ fF}/\mu\text{m}$ of gate width. Further, the effective value of drain-body capacitances are $C_{dbn} = 10 \text{ fF}$ and $C_{dbp} = 15 \text{ fF}$. The wiring capacitance $C_w = 5 \text{ fF}$. Find t_{PHL} , t_{PLH} and t_P .

SOLUTION

First, we determine the value of the equivalent capacitance C using Eq. (13.12),

$$C = 2C_{gd1} + 2C_{gd2} + C_{db1} + C_{db2} + C_{g3} + C_{g4} + C_w$$

where

$$C_{gd1} = 0.5 \times W_n = 0.5 \times 4 = 2 \text{ fF}$$

$$C_{gd2} = 0.5 \times W_p = 0.5 \times 10 = 5 \text{ fF}$$

$$C_{db1} = 10 \text{ fF}$$

$$C_{db2} = 15 \text{ fF}$$

$$C_{g3} = W_n LC_{ox} = 4 \times 2 \times 0.9 = 7.2 \text{ fF}$$

$$C_{g4} = W_p LC_{ox} = 10 \times 2 \times 0.9 = 18 \text{ fF}$$

$$C_w = 5 \text{ fF}$$

Thus,

$$C = 2 \times 2 + 2 \times 5 + 10 + 15 + 7.2 + 18 + 5 = 69.2 \text{ fF}$$

Next, although we can use the formula in Eq. (13.18) to determine t_{PHL} , we shall take an alternative route. Specifically, we shall consider the discharge of C through Q_N and determine the average discharge current using Eqs. (13.14) through (13.16):

$$\begin{aligned} i_{DN}(0) &= \frac{1}{2} k'_n \left(\frac{W}{L} \right)_n (V_{DD} - V_t)^2 \\ &= \frac{1}{2} \times 50 \left(\frac{4}{2} \right) (5 - 1)^2 = 800 \mu\text{A} \end{aligned}$$

$$\begin{aligned} i_{DN}(t_{PHL}) &= k'_n \left(\frac{W}{L} \right)_n \left[(V_{DD} - V_t) \frac{V_{DD}}{2} - \frac{1}{2} \left(\frac{V_{DD}}{2} \right)^2 \right] \\ &= 50 \times \frac{4}{2} \left[(5 - 1) \frac{5}{2} - \frac{1}{2} \left(\frac{5}{2} \right)^2 \right] \\ &= 687.5 \mu\text{A} \end{aligned}$$

Thus

$$i_{DN}|_{av} = \frac{800 + 687.5}{2} = 744 \mu\text{A}$$

and

$$t_{PHL} = \frac{C(V_{DD}/2)}{i_{DN}|_{av}} = \frac{69.2 \times 10^{-15} \times 2.5}{744 \times 10^{-6}} = 233 \text{ ps}$$

Since $\frac{W_p}{W_n} = 2.5 = \frac{\mu_n}{\mu_p}$, the inverter is matched and will have equal propagation delays in both directions, thus

$$t_{PLH} = 233 \text{ ps}$$

$$t_p = \frac{1}{2}(t_{PHL} + t_{PLH}) = 233 \text{ ps}$$

Finally, we note that an identical value would have been obtained had we used the formula in Eq. (13.18) and a slightly lower value (221 ps) would have resulted by using the formula in Eq. (13.13).

Exercises

13.1 Consider the inverter specified in Example 13.1 when loaded with an additional 0.1-pF capacitance. What will the propagation delay become?

Ans. 570 ps

13.2 In an attempt to decrease the area of the inverter in Example 13.1, $(W/L)_p$ is made equal to $(W/L)_n$. What is the percentage reduction in area achieved? Find the new values of C , t_{PHL} , t_{PLH} , and t_p . Assume that C_{dbp} reduces to 10 fF.

Ans. 43%; 47.4 fF; 159 ps; 398 ps; 279 ps

13.3 For the inverter of Example 13.1, find the dynamic power dissipation when clocked at a 100-MHz rate.

Ans. 173 μ W

13.3 CMOS LOGIC-GATE CIRCUITS

In this section, we build on our knowledge of inverter design and consider the design of CMOS circuits that realize combinational logic functions. In combinational circuits, the output at any time is a function only of the values of input signals at that time. Thus, these circuits do not have memory and do not employ feedback. Combinational logic circuits are used in large quantities in a multitude of applications; indeed, every digital system contains large numbers of combinational logic circuits.

Basic Structure

A CMOS logic circuit is in effect an extension, or a generalization, of the CMOS inverter: The inverter consists of an NMOS pull-down transistor, and a PMOS pull-up transistor, operated by the input voltage in a complementary fashion. The CMOS logic gate consists of two networks: the pull-down network (PDN) constructed of NMOS transistors, and the pull-up network (PUN) constructed of PMOS transistors (see Fig. 13.8). The two networks are operated by the input variables, in a complementary fashion. Thus, for the three-input

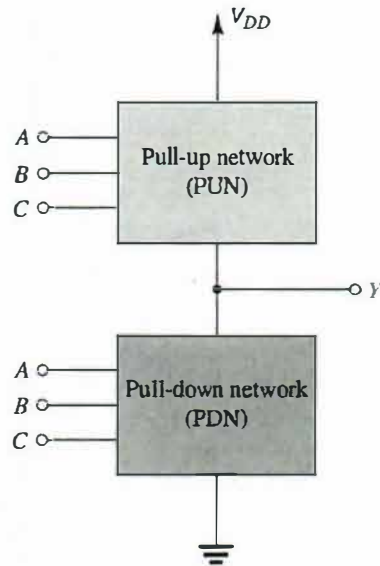


Fig. 13.8 Representation of a three-input CMOS logic gate. The PUN comprises PMOS transistors, and the PDN comprises NMOS transistors.

gate represented in Fig. 13.8, the PDN will conduct for all input combinations that require a low output ($Y = 0$) and will then pull the output node down to ground, causing a zero voltage to appear at the output, $v_Y = 0$. Simultaneously, the PUN will be off, and no direct dc path will exist between V_{DD} and ground. On the other hand, all input combinations that call for a high output ($Y = 1$) will cause the PUN to conduct, and the PUN will then pull the output node up to V_{DD} , establishing an output voltage $v_Y = V_{DD}$. Simultaneously, the PDN will be cut off, and again, no dc current path between V_{DD} and ground will exist in the circuit.

Now since the PDN comprises NMOS transistors, and since the NMOS transistor conducts when the signal at its gate is high, the PDN is activated (that is, conducts) when the inputs are high. In a dual manner, the PUN comprises PMOS transistors, and the PMOS transistor conducts when the input signal at its gate is low, thus the PUN is activated when the inputs are low.

Each of the PDN and the PUN utilizes devices in parallel to form an OR function, and devices in series to form an AND function. Figure 13.9 shows examples of PDNs. For the circuit in Fig. 13.9(a), we observe that Q_A will conduct when A is high ($v_A = V_{DD}$) and will then pull the output node down to ground ($v_Y = 0$ V, $Y = 0$). Similarly, Q_B conducts and pulls Y down when B is high. Thus Y will be low when A is high *or* B is high, which can be expressed as

$$\bar{Y} = A + B$$

or equivalently

$$Y = \overline{A + B}$$

The PDN in Fig. 13.9(b) will conduct only when A and B are both high simultaneously. Thus Y will be low when A is high *and* B is high,

$$\bar{Y} = AB$$

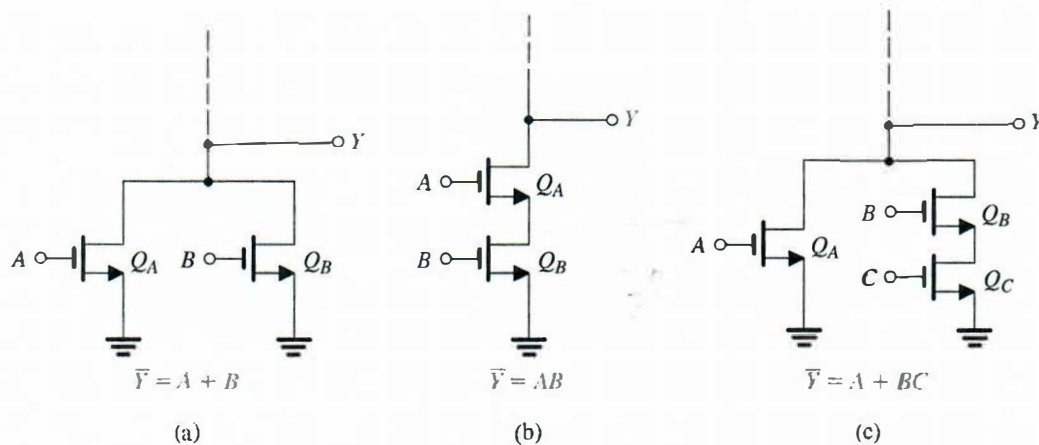


Fig. 13.9 Examples of pull-down networks.

or equivalently

$$Y = \overline{AB}$$

As a final example, the PDN in Fig. 13.9(c) will conduct and cause Y to be 0 when A is high *or* when B and C are both high, thus

$$\overline{Y} = A + BC$$

or equivalently

$$Y = \overline{A + BC}$$

Next consider the PUN examples shown in Fig. 13.10. The PUN in Fig. 13.10(a) will conduct and pull Y up to V_{DD} ($Y = 1$) when A is low *or* B is low, thus

$$Y = \overline{A} + \overline{B}$$

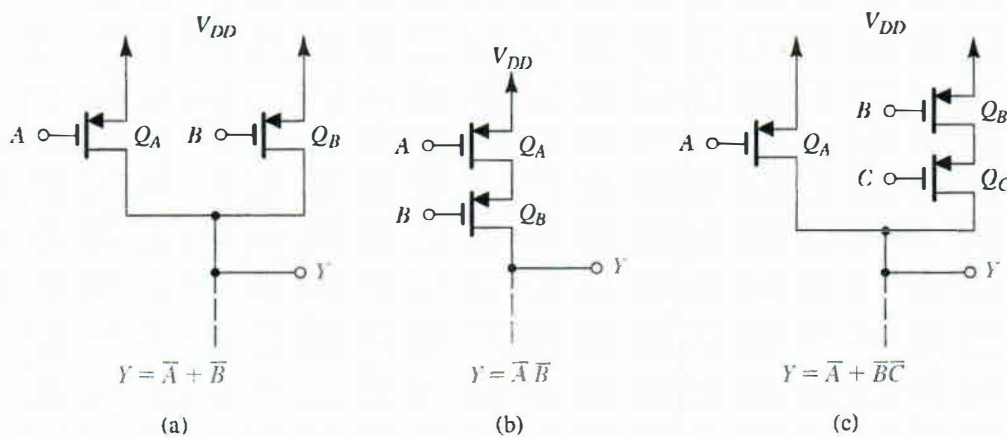


Fig. 13.10 Examples of pull-up networks.

The PUN in Fig. 13.10(b) will conduct and produce a high output ($v_Y = V_{DD}$, $Y = 1$) only when A and B are both low, thus

$$Y = \overline{A} \overline{B}$$

Finally, the PUN in Fig. 13.10(c) will conduct and cause Y to be high (logic 1) if A is low or if B and C are both low, thus

$$Y = \overline{A} + \overline{B} \overline{C}$$

Having developed an understanding and an appreciation of the structure and operation of PDNs and PUNs, we now consider complete CMOS gates. Before doing so, however, we wish to introduce alternative circuit symbols for MOS transistors, that are almost universally used by digital-circuit designers. Figure 13.11 shows our usual symbols and the corresponding new symbols. Observe that the symbol for the PMOS transistor with a circle at the gate terminal is intended to indicate that the signal at the gate has to be low for the device to be activated (i.e., to conduct). Thus in terms of logic-circuit terminology, the gate terminal of the PMOS transistor is an *active low* input. Besides indicating this property of PMOS devices, the new symbols omit any indication of which of the device terminals is the source and which is the drain. This should cause no difficulty at this stage of our study; simply remember that for an NMOS transistor the drain is the terminal that is at the higher voltage (current flows from drain to source), and for a PMOS transistor the source is the terminal that is at the higher voltage (current flows from source to drain). In order to be consistent with the literature, we shall henceforth use these modified symbols for MOS transistors in logic applications, except in locations where our usual symbols help in understanding circuit operation.

The Two-Input NOR Gate

We first consider the CMOS gate that realizes the two-input NOR function

$$Y = \overline{A + B} = \overline{A} \overline{B} \quad (13.21)$$

We see that Y is to be low (PDN conducting) when A is high or B is high. Thus the PDN consists of two parallel NMOS devices with A and B as inputs [that is, the circuit in Fig. 13.9(a)]. For the PUN, we note from the second expression in Eq. (13.21) that Y is to be

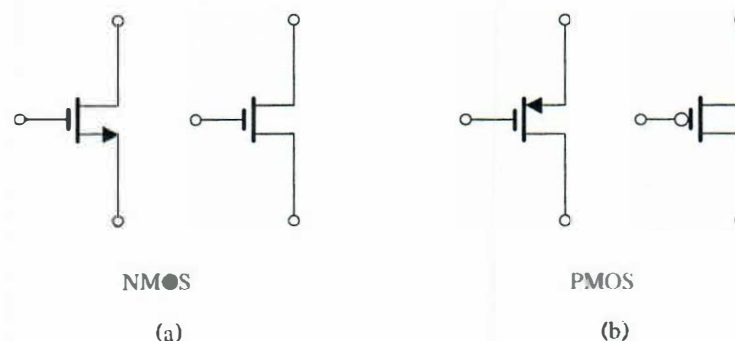


Fig. 13.11 Usual and alternative circuit symbols for MOSFETs.

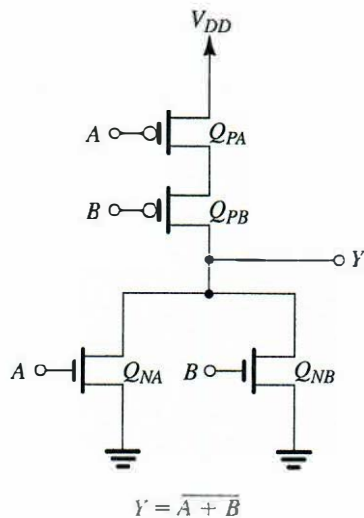


Fig. 13.12 A two-input CMOS NOR gate.

high when A and B are both low. Thus the PUN consists of two series PMOS devices with A and B as the inputs [that is, the circuit in Fig. 13.10(b)]. Putting the PDN and the PUN together gives the CMOS NOR gate shown in Fig. 13.12. Note that extension to a higher number of inputs is straightforward: For each additional input, an NMOS transistor is added in parallel with Q_{NA} and Q_{NB} , and a PMOS transistor is added in series with Q_{PA} and Q_{PB} .

The Two-Input NAND Gate

The two-input NAND function is described by the Boolean expression

$$Y = \overline{AB} = \bar{A} + \bar{B} \quad (13.22)$$

To synthesize the PDN, we consider the input combinations that require Y to be low: There is only one such combination, namely, A and B both high. Thus, the PDN simply comprises two NMOS transistors in series [such as the circuit in Fig. 13.9(b)]. To synthesize the PUN, we consider the input combinations that result in Y being high. These are found from the second expression in Eq. (13.22) as A low or B low. Thus, the PUN consists of two parallel PMOS transistors with A and B applied to their gates [such as the circuit in Fig. 13.10(a)]. Putting the PDN and PUN together results in the CMOS NAND gate implementation shown in Fig. 13.13. Note that extension to a higher number of inputs is straightforward: For each additional input, we add an NMOS transistor in series with Q_{NA} and Q_{NB} , and a PMOS transistor in parallel with Q_{PA} and Q_{PB} .

A Complex Gate

Consider next the more complex logic function

$$Y = \overline{A(B + CD)} \quad (13.23)$$

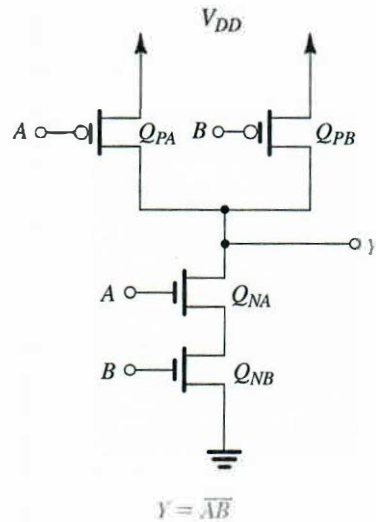


Fig. 13.13 A two-input CMOS NAND gate.

Since $\bar{Y} = A(B + CD)$, we see that Y should be low for A high and simultaneously either B high or C and D both high, from which the PDN is directly obtained. To obtain the PUN, we need to express Y in terms of the complemented variables. We do this through repeated application of DeMorgan's law, as follows:

$$\begin{aligned}
 Y &= \overline{A(B + CD)} \\
 &= \bar{A} + \overline{B + CD} \\
 &= \bar{A} + \bar{B} \bar{CD} \\
 &= \bar{A} + \bar{B}(\bar{C} + \bar{D})
 \end{aligned} \tag{13.24}$$

Thus, Y is high for A low or B low and either C or D low. The corresponding complete CMOS circuit will be as shown in Fig. 13.14.

Obtaining the PUN from the PDN and Vice Versa

From the CMOS gate circuits considered thus far (e.g., that in Fig. 13.14), we observe that the PDN and the PUN are dual networks: where a series branch exists in one, a parallel branch exists in the other. Thus, we can obtain one from the other, a process that can be simpler than having to synthesize each separately from the Boolean expression of the function. For instance, in the circuit of Fig. 13.14, we found it relatively easy to obtain the PDN, simply because we already had $\bar{Y}$ in terms of the uncomplemented inputs. On the other hand, to obtain the PUN, we had to manipulate the given Boolean expression so as to express Y as a function of the complemented variables, the form convenient for synthesizing PUNs. Alternatively, we could have used this duality property to obtain the PUN from the PDN. The reader is urged to refer to Fig. 13.14 to convince herself that this is indeed possible.

It should, however, be mentioned that there are occasions when it is not easy to obtain one of the two networks from the other using the duality property. For such cases, one has

$$Y = A(B+C)$$

(13.23)

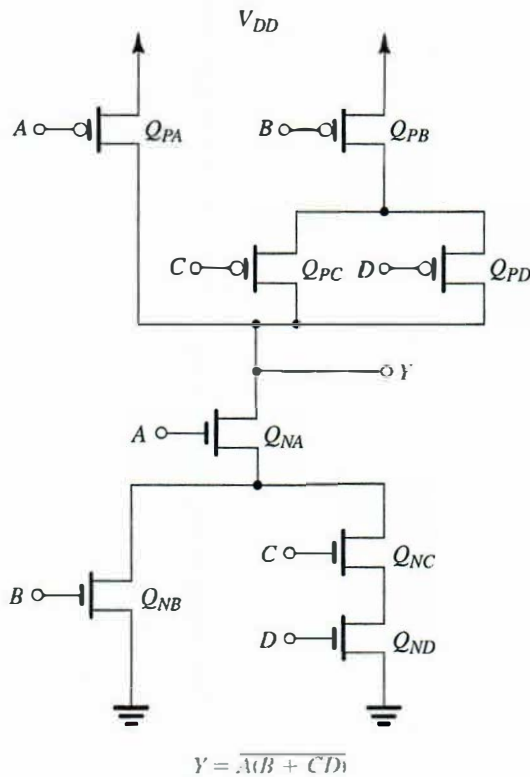


Fig. 13.14 CMOS realization of a complex gate.

to resort to a more rigorous process, which is beyond the scope of this book [see Kang and Leblebici (1996)].

The Exclusive-OR Function

An important function that often arises in logic design is the Exclusive-OR function,

$$Y = A\bar{B} + \bar{A}B \quad (13.25)$$

We observe that since Y (rather than $\bar{Y}$) is given, it is easier to synthesize the PUN. We note, however, that unfortunately Y is not a function of the complemented variables only (as we would like it to be). Thus, we will need additional inverters. The PUN obtained directly from Eq. (13.25) is shown in Fig. 13.15(a). Note that the Q_1, Q_2 branch realizes the first term ($A\bar{B}$), whereas the Q_3, Q_4 branch realizes the second term ($\bar{A}B$). Note also the need for two additional inverters to generate $\bar{A}$ and $\bar{B}$.

As for synthesizing the PDN, we can obtain it as the dual network of the PUN in Fig. 13.15(a). Alternatively, we can develop an expression for $\bar{Y}$ and use it to synthesize the PDN. Leaving the first approach for the reader to do as an exercise, we shall utilize the direct synthesis approach. DeMorgan's law can be applied to the expression in Eq. (13.25) to obtain $\bar{Y}$ as

$$\bar{Y} = AB + \bar{A}\bar{B} \quad (13.26)$$

$$\bar{Y} = \overline{A(B+C)} = \bar{A} + \bar{B}(\bar{C} + \bar{D}) \quad (13.25)$$

$$Y = A(B+C)$$

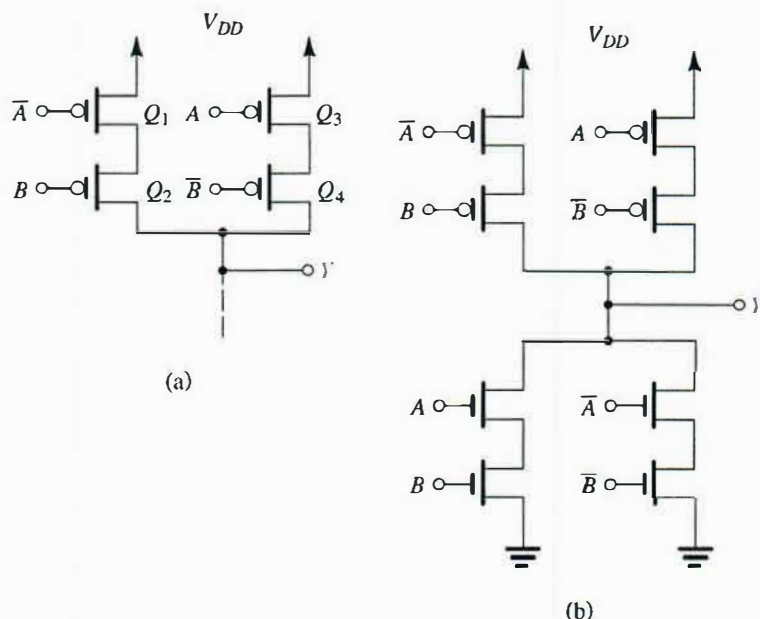


Fig. 13.15 Realization of the Exclusive-OR (XOR) function: (a) The PUN synthesized directly from the expression in Eq. (13.25). (b) The complete XOR realization utilizing the PUN in (a) and a PDN that is synthesized directly from the expression in Eq. (13.26). Note that two inverters (not shown) are needed to generate the complemented variables. Also note that in this XOR realization, the PDN and the PUN are not dual networks; however, a realization based on dual networks is possible (see Problem 13.27).

The corresponding PDN will be as in Fig. 13.15(b), which shows the CMOS realization of the Exclusive-OR function except for the two additional inverters. Note that the Exclusive-OR requires 12 transistors for its realization, a rather complex network. We shall show a simpler realization of the Exclusive-OR in Section 13.5 employing a different form of CMOS logic.

Another interesting observation follows from the circuit in Fig. 13.15(b). The PDN and the PUN here are *not* dual networks. Indeed, duality of the PDN and the PUN is not a necessary condition. Thus, although a dual of PDN (or PUN) can always be used for PUN (or PDN), the two networks are not necessarily duals.

Summary of the Synthesis Method

1. The PDN can be most directly synthesized by expressing $\bar{Y}$ as a function of the *uncomplemented* variables. If complemented variables appear in this expression, additional inverters will be required to generate them.
2. The PUN can be most directly synthesized by expressing Y as a function of the *complemented* variables and then applying the uncomplemented variables to the

gates of the PMOS transistors. If uncomplemented variables appear in the expression, additional inverters will be needed.

3. The PDN can be obtained from the PUN (and vice versa) using the duality property.

Transistor Sizing

Once a CMOS gate circuit has been generated, the only significant step remaining in the design is to decide on W/L ratios for all devices. This selection is usually done so as to provide the gate with current-driving capability in both directions equal to that of the basic inverter. The reader will recall from the previous section that for the basic inverter design, we denoted $(W/L)_n = n$ and $(W/L)_p = p$, where n is usually 1.5 to 2 and, for a matched design, $p = (\mu_n/\mu_p)n$. Thus, we wish to select individual W/L ratios for all transistors in a logic gate so that the PDN should be able to provide a capacitor discharge current *at least* equal to that of an NMOS transistor with $W/L = n$, and the PUN should be able to provide a charging current *at least* equal to that of a PMOS transistor with $W/L = p$. This will guarantee a *worst-case* gate delay equal to that of the basic inverter.²

In the preceding description, the idea of "worst case" should be emphasized. It means that in deciding on device sizing, we should find the input combinations that result in the lowest output current and then choose sizes that will make this current equal to that of the basic inverter. Before we consider examples, we need to address the issue of determining the current-driving capability of a circuit consisting of a number of MOS devices. In other words, we need to find the *equivalent W/L ratio* of a network of MOS transistors. Toward that end, we consider the parallel and series connection of MOSFETs and find the equivalent W/L ratios.

The derivation of the equivalent W/L ratio is based on the fact that the on resistance of a MOSFET is inversely proportional to W/L . Thus if a number of MOSFETs having ratios of $(W/L)_1, (W/L)_2, \dots$ are connected in series, the equivalent series resistance obtained by adding the on resistances will be

$$\begin{aligned} R_{\text{series}} &= r_{DS1} + r_{DS2} + \dots \\ &= \frac{\text{constant}}{(W/L)_1} + \frac{\text{constant}}{(W/L)_2} + \dots \\ &= \text{constant} \left[\frac{1}{(W/L)_1} + \frac{1}{(W/L)_2} + \dots \right] \\ &= \frac{\text{constant}}{(W/L)_{eq}} \end{aligned}$$

resulting in the following expression for $(W/L)_{eq}$ for transistors connected in series:

² This statement assumes that the total effective capacitance C of the logic gate is the same as that of the inverter. In actual practice, the value of C will be larger for a gate, especially as the fan-in is increased.

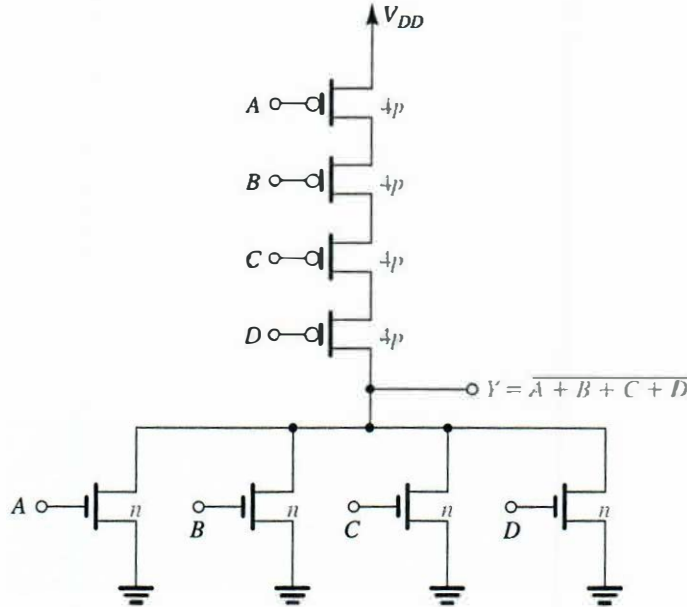


Fig. 13.16 Proper transistor sizing for a four-input NOR gate. Note that n and p denote the (W/L) ratios of Q_N and Q_P , respectively, of the basic inverter.

$$(W/L)_{eq} = \frac{1}{\frac{1}{(W/L)_1} + \frac{1}{(W/L)_2} + \dots} \quad (13.27)$$

Similarly, we can show that the parallel connection of transistors with W/L ratios of $(W/L)_1, (W/L)_2, \dots$, results in an equivalent W/L of

$$(W/L)_{eq} = (W/L)_1 + (W/L)_2 + \dots, \quad (13.28)$$

As an example, two identical MOS transistors with individual W/L ratios of 4 result in an equivalent W/L of 2 when connected in series and of 8 when connected in parallel.

As an example of proper sizing, consider the four-input NOR in Fig. 13.16. Here, the worst case (the lowest current) for the PDN is obtained when only one of the NMOS transistors is conducting. We therefore select the W/L of each NMOS transistor to be equal to that of the NMOS transistor of the basic inverter, namely, n . For the PUN, however, the worst-case situation (and indeed the only case) is when all inputs are low and the four series PMOS transistors are conducting. Since the equivalent W/L will be one quarter of that of each PMOS device, we should select the W/L ratio of each PMOS transistor to be four times that of Q_P of the basic inverter, that is, $4p$.

As another example, we show in Fig. 13.17 the proper sizing for a four-input NAND gate. Comparison of the NAND and NOR gates in Figs. 13.16 and 13.17 indicates that because p is usually two to three times n , the NOR gate will require much greater area than the NAND gate. For this reason, NAND gates are generally preferred for implementing combinational logic functions in CMOS.

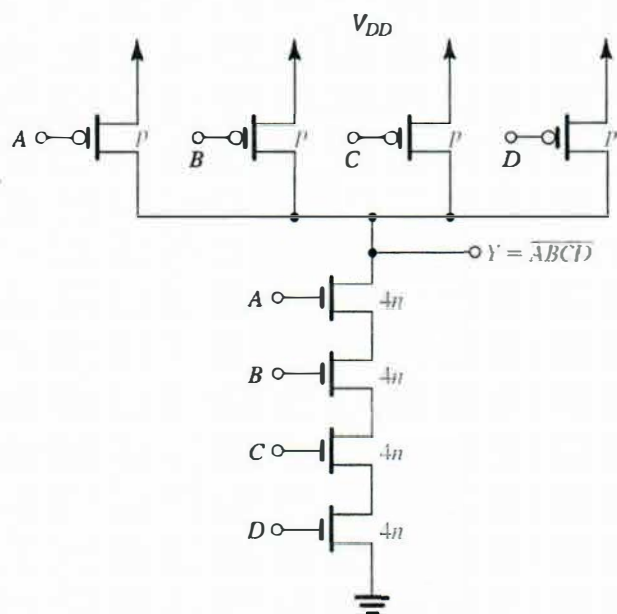


Fig. 13.17 Proper transistor sizing for a four-input NAND gate. Note that n and p denote the (W/L) ratios of Q_N and Q_P , respectively, of the basic inverter.

EXAMPLE 13.2

Provide transistor W/L ratios for the logic circuit shown in Fig. 13.18. Assume that for the basic inverter $n = 2$ and $p = 5$ and that the channel length is $2\ \mu\text{m}$.

SOLUTION

Refer to Fig. 13.18, and consider the PDN first. We note that the worst case occurs when Q_{NB} is on and either Q_{NC} or Q_{ND} is on. That is, in the worst case, we have two transistors in series. Therefore, we select each of Q_{NB} , Q_{NC} , and Q_{ND} to be of twice the width of the n -channel device in the basic inverter, thus

$$Q_{NB}: W/L = 2n = 4 = 8/2$$

$$Q_{NC}: W/L = 2n = 4 = 8/2$$

$$Q_{ND}: W/L = 2n = 4 = 8/2$$

For transistor Q_{NA} , select W/L to be equal to that of the n -channel device in the basic inverter:

$$Q_{NA}: W/L = n = 2 = 4/2$$

Next, consider the PUN. Here, we see that in the worst case, we have three transistors in series: Q_{PA} , Q_{PC} , and Q_{PD} . Therefore, we select the W/L ratio of each of these to be

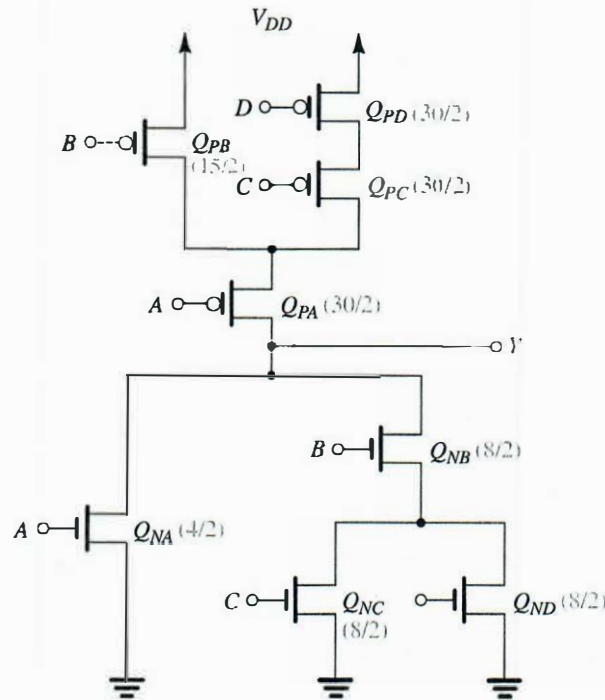


Fig. 13.18 Circuit for Example 13.2.

three times that of Q_P in the basic inverter, that is, $3p$, thus

$$Q_{NA}: W/L = 3p = 15 = 30/2$$

$$Q_{NC}: W/L = 3p = 15 = 30/2$$

$$Q_{ND}: W/L = 3p = 15 = 30/2$$

Finally, the W/L ratio for Q_{PB} should be selected so that the equivalent W/L of the series connection of Q_{PB} and Q_{PA} should be equal to p . It follows that for Q_{PB} the ratio should be $1.5p$,

$$Q_{PB}: W/L = 1.5p = 7.5 = 15/2$$

Figure 13.18 shows the circuit with the transistor sizes indicated.

Effects of Fan-In and Fan-Out on Propagation Delay

Each additional input to a CMOS gate requires two additional transistors, one NMOS and one PMOS. This is in contrast to other forms of MOS logic where each additional input requires only one additional transistor. The additional transistor in CMOS not only increases the chip area but also increases the total effective capacitance per gate and in turn increases the propagation delay. The size-scaling method described earlier compensates for some but not all of the increase in t_p . Specifically, by increasing device size, we are able to preserve the current-driving capability. However, the capacitance C increases because of both the increased number of inputs and the increase in device size. Thus t_p will still increase with fan-in, a fact that imposes a practical limit on the fan-in of, say, the NAND gate to about

4. If a higher number of inputs is required, then “clever” logic design should be adopted so as to realize the given Boolean function with gates of no more than four inputs. This would usually mean an increase in the number of cascaded stages and thus an increase in delay. However, such an increase in delay can be less than the increase due to the large fan-in (see Problem 13.36).

An increase in gate fan-out adds directly to its load capacitance and thus increases its propagation delay.

In conclusion, although CMOS has many advantages, it does suffer from increased circuit complexity when the fan-in and fan-out are increased, and the corresponding effects of this complexity on both chip area and propagation delay. In the following two sections, we shall study some simplified forms of CMOS logic that attempt to solve this complexity problem, although at the expense of foregoing some of the advantages of basic CMOS.

Exercises

13.4 For a process technology with $L = 1.2 \mu\text{m}$, $n = 1.5$, $p = 4.5$, give the sizes of all transistors in (a) four-input NOR and (b) four-input NAND. Also give the relative areas of the two gates.

Ans. (a) NMOS devices: $W/L = 1.8/1.2$, PMOS devices: $21.6/1.2$
 (b) NMOS devices: $W/L = 7.2/1.2$, PMOS devices: $5.4/1.2$; NOR area/NAND area = 1.86

13.5 For the scaled NAND gate in the preceding exercise, what is the ratio of the maximum to minimum current available to (a) charge a load capacitance? (b) discharge a load capacitance?

Ans. (a) 4; (b) 1

13.4 PSEUDO-NMOS LOGIC CIRCUITS

As explained in the previous section, despite its many great advantages, CMOS suffers from increased area, and correspondingly increased capacitance and delay, as the logic gates become more complex. For this reason, digital integrated-circuit designers have been searching for forms of CMOS logic circuits that can be used to supplement the complementary-type circuits studied in the two previous sections. These forms are not intended to replace complementary CMOS but rather to be used in special applications for special purposes. We shall examine two such CMOS logic styles in this and the following section.

The Pseudo-NMOS Inverter

Figure 13.19 shows a modified form of the CMOS inverter. Here, only Q_N is driven by the input voltage while the gate of Q_P is grounded, and Q_P acts as an active load for Q_N . Even before we examine the operation of this circuit in detail, an advantage over complementary CMOS is immediately obvious: Each input must be connected to the gate of only one transistor or, alternatively, only one additional transistor (an NMOS) will be needed for each additional gate input. Thus the area and delay penalties arising from increased fan-in

APPENDIX A

VLSI Fabrication Technology

INTRODUCTION

The purpose of this appendix is to familiarize the reader with VLSI (very large scale integrated-circuits) fabrication technology. Brief explanations of the standard VLSI processing steps are given. The characteristics of devices available in **CMOS** and **BiCMOS** fabrication technologies are also presented. This is done to aid the reader in understanding those aspects of IC (integrated-circuit) design that are distinct from discrete-circuit design. In order to take proper advantage of the economics of integrated circuits, designers have had to overcome some serious device limitations (such as poor device tolerances) while exploiting device advantages (such as good component matching). An understanding of device characteristics is therefore essential in designing good custom ICs or ASICs (application-specific ICs), and also helps when applying commercially available ICs to system design.

This appendix will consider only silicon (Si) technology. Although germanium (Ge) and gallium arsenic (GaAs) are also used to implement VLSI chips, silicon is still the most popular material and will remain so for some time. Silicon is an abundant element and occurs naturally in the form of sand. It can be refined using simple purification and crystal growth techniques. It also exhibits suitable physical properties for fabricating active devices with good electrical characteristics. In addition, silicon can be easily oxidized to form an excellent insulator, SiO_2 (glass). This native oxide is useful for constructing capacitors and MOSFETs. It also serves as a diffusion barrier that masks against unwanted impurities from diffusing into the high purity silicon material. This masking property allows selective alteration of the electrical properties in the silicon; thus active and passive elements can be built on the same piece of material (substrate). The components can then be interconnected to form a monolithic IC.

IC FABRICATION STEPS

The basic IC fabrication steps will be described in the following sections. Some of these steps may be carried out several times, in different combinations and operating conditions during a complete fabrication run.

Wafer Preparation

The starting material for modern integrated circuits is very high purity silicon. The material is grown as a single crystal ingot. It takes the shape of a steel gray solid cylinder 10 to 30 cm in diameter and about 1 meter in length. This crystal is then sawed (like a loaf of bread) to produce wafers 10 to 30 cm

in diameter and 400 to 600 μm thick (a micrometer or micron is a millionth of a meter); see Fig. A.1. The surface of the wafer is then polished to a mirror finish using chemical and mechanical polishing (CMP) techniques. Semiconductor manufacturers usually purchase ready-made silicon wafers from a foundry and rarely start their process in ingot form.

The basic electrical and mechanical properties of the wafer depend on the orientation of the crystal growth, the concentration and the type of impurities present. These variables are strictly controlled during crystal growth. Impurities can be added on purpose to the pure silicon in a process known as doping. This allows controlled alteration of the electrical properties of the silicon, in particular, the resistivity. It is also possible to control the type of carrier used to produce electrical conduction, being either holes (in *p*-type silicon) or electrons (in *n*-type silicon). If a large number of impurity atoms is added, then the silicon is said to be heavily doped. When designating relative doping concentrations on diagrams of semiconductor devices, it is common to use + and - symbols. A heavily doped (low-resistivity) *n*-type silicon wafer would be referred to as *n*+ material, whereas a lightly doped region may be referred to as *n*-. The ability to control the doping of silicon permits the formation of diodes, transistors, and resistors in integrated circuits.

Oxidation

Oxidation refers to the chemical process of silicon reacting with oxygen to form silicon dioxide, SiO_2 . To speed up the reaction, it is necessary to heat the wafers to the 1000°C to 1200°C range. The heating is performed in special high-temperature ultra-clean furnaces. To avoid the introduction of even small quantities of contaminants (which could significantly alter the electrical properties of the silicon), it is necessary to maintain a clean environment for the processing. This is true for all processing steps involved in the fabrication of an integrated circuit. Specially filtered air is circulated in the processing area, and all personnel must wear special lint-free clothing.

The oxygen used in the reaction can be introduced either as a high-purity gas (referred to as a "dry oxide") or as water vapor (forming a "wet oxide"). In general, wet oxidation has a faster growth rate, but dry oxidation gives better electrical characteristics. The thermally grown oxide layer has excellent electrical insulation properties. It has a dielectric constant of about 3.9 and can be used to form excellent capacitors. It also serves as an effective mask against many impurities, allowing the introduction of dopants into the silicon only in regions that are not covered with oxide. This masking property is what permits the convenient fabrication of integrated circuits. Silicon dioxide can also be used to protect the silicon surface from contaminants after the VLSI chip is fabricated.

Silicon dioxide is a thin, transparent film, and the silicon surface is highly reflective. If white light is incident on the oxidized wafer, constructive and destructive interference effects occur in the oxide, causing certain colors to be absorbed strongly. The wavelengths absorbed depend on the thickness of the oxide layer, producing different colors in the different regions of a processed wafer. The colors can be quite vivid and are immediately obvious when a finished wafer is viewed with the naked eye. It

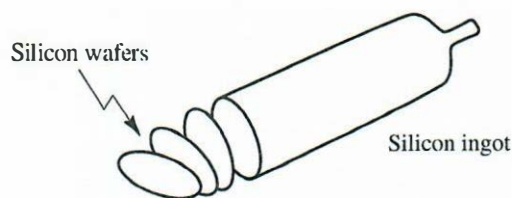


Fig. A.1 Silicon ingot and wafer slices.

should be remembered, though, that the color is due to an optical effect. The oxide layer is transparent, and the silicon underneath it is steel-gray in color.

Diffusion

Diffusion is the process by which atoms move through the crystal lattice. This is very much like a drop of ink dispersed through a glass of water except that it occurs very slowly in solids. In fabrication, this introduces impurity atoms (dopants) into silicon to change its resistivity. The rate at which dopants diffuse in silicon is a strong function of temperature. Diffusion of impurities is usually carried out at high temperatures (1000°C to 1200°C) to obtain the desired doping profile. When the wafer is cooled to room temperature, the impurities are essentially “frozen” in position. The diffusion process is performed in furnaces similar to those used for oxidation. The depth to which the impurities diffuse depends on both the temperature and time allowed.

The most common impurities used as dopants are boron, phosphorus, and arsenic. Boron is a *p*-type dopant, whereas phosphorus and arsenic are *n*-type dopants. These dopants can be effectively masked by thin silicon dioxide layers. By diffusing boron into an *n*-type substrate, a *pn* junction is formed (diode). If the doping concentration is heavy, the diffused layer can also be used as a conductor.

Ion Implantation

Ion implantation is another method used to introduce impurities into silicon. An ion implanter produces ions of the desired impurity, accelerates them by an electric field, and allows them to strike the silicon surface. The ions become embedded in the silicon. The depth of penetration is related to the energy of the ion beam, which can be controlled by the accelerating-field voltage. The quantity of ions implanted can be controlled by varying the beam current (flow of ions). Since both voltage and current can be accurately measured and controlled, ion implantation results in much more accurate and reproducible impurity profiles than can be obtained by diffusion. In addition, ion implantation can be performed at room temperature. Ion implantation normally is used when accurate control of the dopant is essential for device operation.

Chemical Vapor Deposition

Chemical vapor deposition (CVD) is a process by which gases or vapors are chemically reacted, leading to the formation of a solid on a substrate. CVD can be used to deposit silicon dioxide on a silicon substrate. For instance, if silane gas and oxygen are mixed above a silicon substrate, silicon dioxide deposits as a solid on the silicon. The oxide layer formed is not as good as a thermally grown oxide, but it is good enough to act as an electrical insulator. The advantage of a CVD layer is that the oxide deposits at a faster rate and a lower temperature (below 500°C).

If silane gas alone is used, then a silicon layer deposits on the wafer. If the reaction temperature is high enough (above 1000°C), then the layer is deposited as a crystalline layer (assuming the substrate is crystalline silicon). This is because the atoms have enough energy to align themselves in the proper crystal direction. Such a layer is said to be an **epitaxial** layer, and the deposition process is referred to as **epitaxy** instead of CVD. At lower temperatures, or if the substrate surface is not single-crystal silicon, the atoms are not all aligned along the same crystal direction. Such a layer is called polycrystalline silicon since it consists of many small crystals of silicon aligned in various directions. Such layers are normally doped very heavily to form a high conductivity region that can be used for interconnecting devices.

Metallization

The purpose of metallization is to interconnect the various components of the integrated circuit (transistors, capacitors, etc.) to form the desired circuit. Metallization involves the deposition of a metal (aluminum) over the entire surface of the silicon. The required interconnection pattern is then selectively etched. The aluminum is deposited by heating it in vacuum until it vaporizes. The vapors then contact the silicon surface and condense to form a solid aluminum layer.

Photolithography

The surface geometry of the various integrated-circuit components is defined photographically: The silicon surface is coated with a photosensitive layer called photoresist. When exposed to light through a master pattern on a photographic plate, the photoresist will become softened (for positive photoresist). The exposed layer can then be removed using a chemical developer, causing the mask pattern to appear on the wafer. Very fine surface geometries can be reproduced accurately by this technique. The resulting layer is not attacked by the chemical etchants used for silicon dioxide or aluminum and so forms an effective mask. This allows "windows" to be etched in the oxide layer in preparation for subsequent diffusion processes. This process is used to define transistor regions and to isolate one transistor from another.

In high resolution processes, deep UV light is used to expose the photoresist. Alternatively, an electron beam can be used to raster scan or "write" the photoresist directly without the use of a photographic plate.

Packaging

A finished silicon wafer may contain more than several hundreds of finished circuits or chips. Each chip may contain between 10^4 and 10^9 transistors, and is rectangular in shape, typically, between 1 and 10 mm on each side. The circuits are first tested electrically (while still in wafer form) using an automatic probing station. Bad circuits are marked for later identification. The circuits are then separated from each other (by dicing), and the good circuits (dies) are mounted in packages (headers). Fine gold wires are traditionally used to interconnect the pins of the package to the metallization pattern on the die. Finally, the package is sealed under vacuum or in an inert atmosphere. Figure A.2 shows a popular IC package.

VLSI PROCESSES

Integrated-circuit fabrication technology was originally dominated by bipolar technology. By the late 1970s, MOS (metal oxide semiconductor) technology became more promising for VLSI implementation with higher packing density and lower power consumption. Since the early 1980s, CMOS (complementary MOS) technology has almost completely dominated the VLSI scene, leaving bipolar technology to fill specialized functions such as high-speed analog and RF circuits. CMOS technologies continue to evolve, and in the late 1980s, the incorporation of bipolar devices has led to the emergence of high-performance BiCMOS (bipolar-CMOS) fabrication processes that provided the best of both tech-

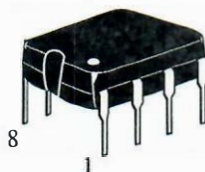


Fig. A.2 A typical 8-pin plastic dual-in-line IC package.

nologies. However, BiCMOS processes are often very complicated and costly since it require upwards of 15 to 20 masking levels per implementation—standard CMOS processes by comparison require only 10 to 12 masking levels. Furthermore, the performance gap is becoming smaller each year as the performance of CMOS continues to improve with finer lithography resolution. For these reasons, CMOS technology will continue to remain the workhorse of most VLSI implementations.

In this section, we will examine a typical CMOS process flow, the performance of the available components, and the inclusion of bipolar devices to form a BiCMOS process.

***n*-Well CMOS Process**

Depending on the choice of starting material (substrate), CMOS processes can be identified as *n*-well, *p*-well, or *twin-well* processes. The last being the most complicated but most flexible in the optimization of both the *n*- and *p*-channel devices.

An *n*-well CMOS process is chosen for discussion since it can be easily extended into a BiCMOS process. The typical process flow is as shown in Fig. A.3. A minimum of seven masking layers are necessary. However, in practice, most CMOS processes will also require additional layers such as *n*- and *p*-guards for better latchup immunity, a second polysilicon layer for capacitors, and possibly multilayer metals for high-density interconnections. The inclusion of these layers would increase the total number of masking layers to more than ten. For simplicity, these layers are not discussed in this appendix.

The *n*-well process begins with the *n*-well diffusion [Fig. A.3(a)]. The *n*-well is required wherever *p*-type MOSFETs are to be placed. A thick silicon dioxide layer is etched to expose the regions for *n*-well diffusion. The unexposed regions will be protected from the phosphorous impurity. Phosphorous is usually used for deep diffusions since it has a high diffusion coefficient and can diffuse faster into the substrate than can arsenic.

The second step is to define the active region (region where transistors are to be placed) using a technique called **local oxidation (LOCOS)**. A silicon nitride (Si_3N_4) layer is deposited and patterned relative to the previous *n*-well regions [Fig. A.3(b)]. The nitride covered regions will not be oxidized. After a long wet oxidation step, thick field oxide will appear in regions between transistors [Fig. A.3(c)]. This thick field oxide is necessary for isolating the transistors. It also allows interconnection layers to be routed on top without inadvertently forming a conduction channel at the silicon surface.

The next step is the formation of the polysilicon gate [Fig. A.3(d)]. This is one of the most critical steps in the CMOS process. The thin oxide layer in the active region is first removed using wet etching followed by the growth of a high-quality thin gate oxide. Current 0.25- and 0.5- μm processes routinely make use of oxide thicknesses as thin as 200 Å (1 angstrom = 10^{-8} cm). A polysilicon layer, usually arsenic doped (*n*-type), is deposited and patterned. The photolithography is most demanding in this step since the finest resolution is required to produce the shortest possible MOS channel length. This is represented by the size of the narrowest polysilicon strip that can be defined.

The polysilicon gate is a self-aligned structure and is preferred to the older type of metal gate structure. A heavy arsenic implant can be used to form the *n*+ source and drain regions of the *n*-MOSFETs. The polysilicon gate also acts as a barrier for this implant to protect the channel region. A layer of photoresist can be used to block the regions where *p*-MOSFETs are to be formed [Fig. A.3(e)]. The thick field oxide stops the implant and prevents *n*+ regions from forming outside the active regions.

A reversed photolithography step can be used to protect the *n*-MOSFETs during the *p*+ boron source and drain implant for the *p*-MOSFETs [Fig. A.3(f)]. Note that in both cases the separation between the source and drain diffusions—channel length—is defined by the polysilicon gate mask alone, hence the self-aligned property.

Before contact holes are opened, a thick layer of CVD oxide is deposited over the entire wafer. A photo-mask is used to define the contact window opening [Fig. A.3(g)] followed by a wet or dry oxide



Si_3N_4

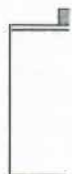
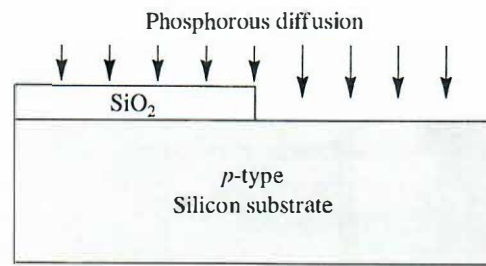
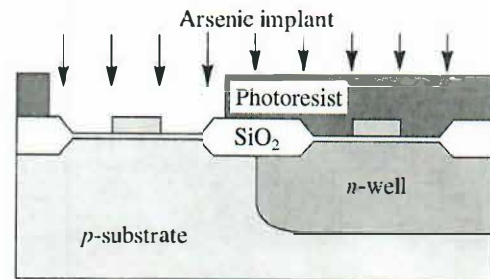
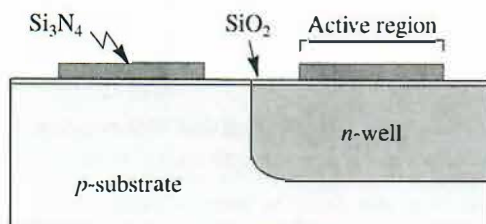
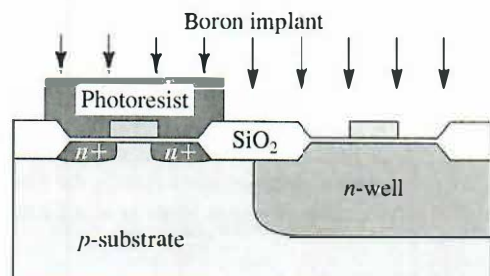
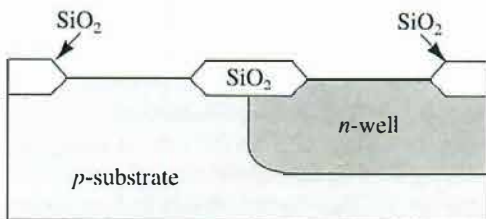


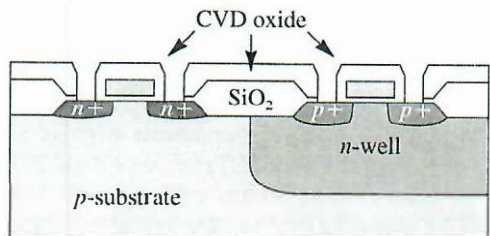
Fig.

(a) Define n -well diffusion (mask 1)(e) n^+ diffusion (mask 4)

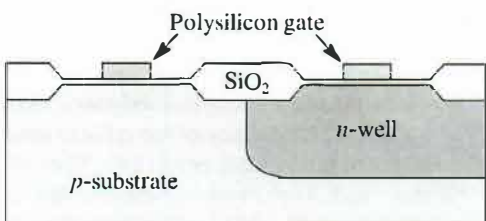
(b) Define active regions (mask 2)

(f) p^+ diffusion (mask 5)

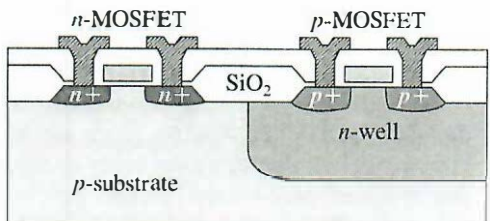
(c) LOCOS oxidation



(g) Contact holes (mask 6)



(d) Polysilicon gate (mask 3)



(h) Metallization (mask 7)

Fig. A.3 A typical n -well CMOS process flow.

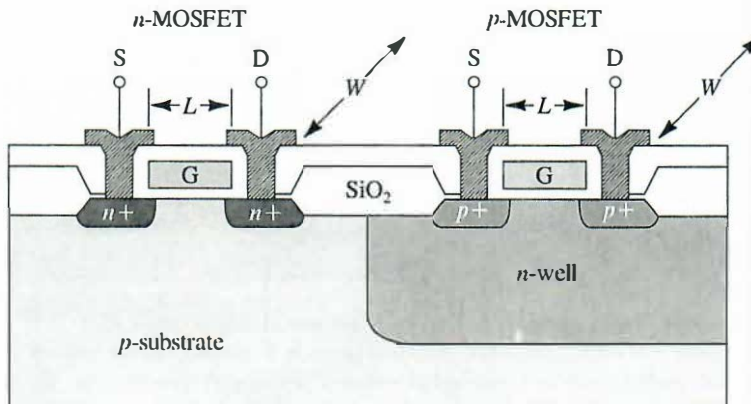


Fig. A.4 Cross-sectional diagram of an n - and p -MOSFET.

etch. A thin aluminum layer is then evaporated or sputtered onto the wafer. A final masking and etching step is used to pattern the interconnection [Fig. A.3(h)].

Not shown in the process flow is the final passivation step prior to packaging and wire bonding. A thick CVD oxide or pyrox glass is usually deposited on the wafer to serve as a protective layer.

CMOS Devices

Besides the obvious n - and p -channel MOSFETs, there are other devices that can be obtained by manipulating the masking layers. These include pn junction diodes, MOS capacitors, and resistors.

MOSFETs

The n -channel MOSFET is preferred to the p -MOSFET. This is because the electron surface mobility is two to three times higher than that for holes, and thus with the same device size (W and L), the n -MOSFET offers higher current drive (or lower on resistance) and higher transconductance.

Integrated MOSFETs are characterized by their threshold voltage and device sizes. Usually, the n - and p -channel devices are designed to have threshold voltages of equal magnitude and remain fixed for a particular process. The transconductance can be adjusted by changing the device surface dimensions (W and L): see Fig. A.4. This feature is not available for bipolar transistors, thus the design of integrated MOSFET circuits is much more straightforward.

Resistors

Resistors in integrated form are not very precise. They can be made from various diffusions, as shown in Fig. A.5. Different diffused regions have different resistivity. The n -well is usually used for medium-value resistors, whereas the $n+$ and $p+$ diffusions are useful for low-value resistors. The actual resistance value can be defined by changing the length and width of diffused regions. The tolerance of the resistor value is very poor (20% to 50%), but the matching of two similar resistor values is quite good (5%). Thus circuit designers should design circuits that exploit resistor matching and should avoid all circuits that require a specific resistor value. Also note that diffused resistors have a significant temperature coefficient.

All diffused resistors are self-isolated by the reverse-biased pn junctions. One serious drawback for these resistors is the fact that they are accompanied by a substantial parasitic junction capacitance, making them not very useful for high-frequency applications.

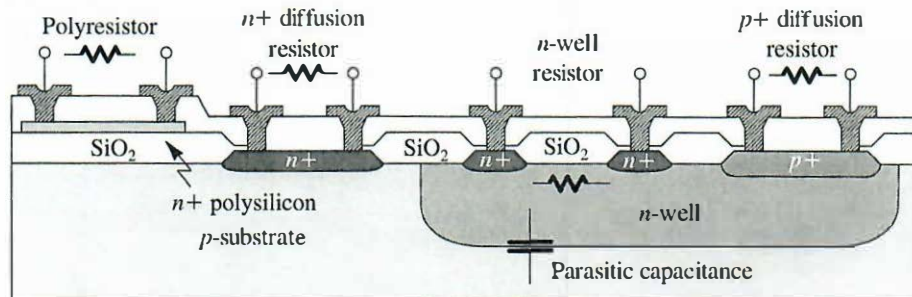


Fig. A.5 Cross sections of various types of resistors available from a typical *n*-well CMOS process.

A more useful resistor can be fabricated using the polysilicon layer placed on top of the thick field oxide. The thin layer provides a better surface area matching and hence more accurate resistor ratios. Furthermore, the polyresistor is physically separated from the substrate and exhibits a much lower parasitic capacitance.

Capacitors

Two types of capacitor structures are available in CMOS processes, MOS and inter-poly capacitors. The cross sections of these structures are shown in Fig. A.6. The MOS gate capacitance, depicted by the center structure, is basically the gate-to-source capacitance of a MOSFET. The capacitance value is dependent on the gate area. The oxide thickness is the same as the gate oxide thickness in the MOSFETs. This capacitor exhibits a large voltage dependence. To eliminate this problem, an additional *n*+ implant is required to form the bottom plate of the capacitors, as shown in the structure on the right. Both of these MOS capacitors are physically in contact with the substrate, resulting in a large parasitic *pn* junction capacitance at the bottom plate.

The inter-poly capacitor exhibits near ideal characteristics but at the expense of the addition of a second poly layer to the CMOS process. Since this capacitor is placed on top of the thick field oxide, parasitic effects are kept to a minimum.

A third and less often used capacitor is the junction capacitor. Any *pn* junction under reverse bias produces a depletion region that acts as a dielectric between the *p*- and the *n*-regions. The capacitance is determined by geometry and doping levels and with a large voltage coefficient. The fact that this capacitor works only with reverse bias voltages makes it less useful.

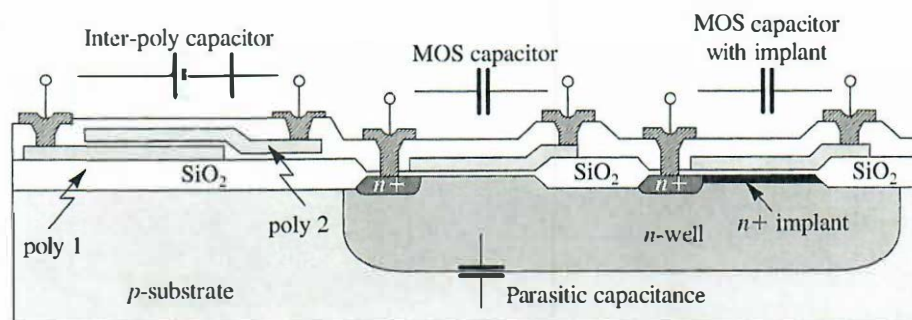


Fig. A.6 Inter-poly and MOS capacitors in an *n*-well CMOS process.

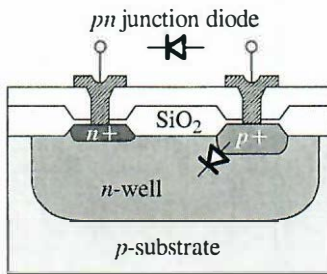


Fig. A.7 pn junction diode in an n -well CMOS process.

For inter-poly and the MOS capacitors, the capacitance values can be controlled to within 1%. Practical capacitance values range from 0.5 pF to a few 10s of pF. The matching between similar-size capacitors can be within 0.1%. This property is extremely useful for designing precision analog CMOS circuits.

pn Junction Diodes

Whenever n -type and p -type diffusion regions are placed next to each other, a pn junction diode results. A useful structure is the n -well diode shown in Fig. A.7. The fact that the diode is fabricated in an n -well provides a high breakdown voltage. These diodes are essential for the input clamping circuits for protection against electrostatic discharge. The diode is also very useful as an on-chip temperature sensor by monitoring the variation of its forward voltage drop.

BiCMOS Process

An npn vertical bipolar transistor can be integrated into the n -well CMOS process with the addition of a p -base diffusion region (see Fig. A.8). The characteristics of this device depend on the base width and the emitter area. The base width is determined by the difference in junction depth between the $n+$ and the p -base diffusions. The emitter area is determined by the junction area of the $n+$ diffusion at the emitter. The n -well serves as the collector for the npn transistor. Typically, the npn transistor has a β in the range of 50 to 100 and a cutoff frequency of greater than 10 GHz.

Normally, an $n+$ buried layer is used to reduce the series resistance of the collector since the n -well has a very high resistivity. However, this would further complicate the process with the introduction of p -type epitaxy and one more masking step. Other variations on the bipolar transistor include poly emitter and self-aligned base contact to minimize parasitic effects.

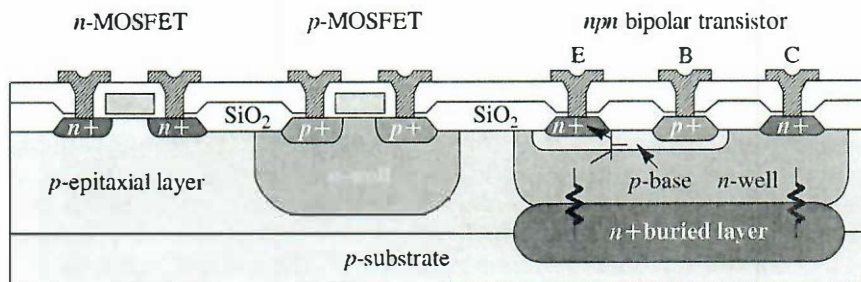


Fig. A.8 Cross-sectional diagram of a BiCMOS process.

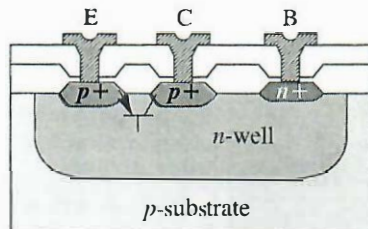


Fig. A.9 A lateral *pnp* transistor.

Lateral *pnp* Transistor

The fact that most BiCMOS processes do not have optimized *pnp* transistors makes circuit design somewhat difficult. However, in noncritical situations, a parasitic lateral *pnp* transistor can be used (Fig. A.9).

In this case, the *n*-well serves as the *n*-base region, with the *p*+ diffusions as the emitter and the collector. The base width is determined by the separation between the two *p*+ diffusions. Since the doping profile is not optimized for the base-collector junctions and because the base width is limited by the minimum photolithographic resolution, the performance of this device is not very good, typically, β of around 10 and with a low cutoff frequency.

p-Base and Pinched-Base Resistors

With the additional *p*-base diffusion in the BiCMOS process, two additional resistor structures are available. The *p*-base diffusion can be used to form a straightforward *p*-base resistor, as shown in Fig. A.10. Since the base region is usually of a relatively low doping level and with a moderate junction depth, it is suitable for medium-value resistors (a few $k\Omega$ s). If large resistor value is required, the pinched-base resistor can be used. In this structure, the *p*-base region is encroached by the *n*+ diffusion, restricting the conduction path. Resistor values in the range of 10 $k\Omega$ to 100 $k\Omega$ can be obtained. As with the diffusion resistors discussed earlier, these resistors exhibit similarly poor tolerance and temperature coefficients.

VLSI LAYOUT

The designed circuit schematic must be transformed into a layout that consists of the geometric representation of the circuit components and interconnections. With the advent of CAD (computer-aided design) tools, many of the conversion steps from schematic to layout can be carried out semi- or fully automatically. However, any good IC designer must have practiced full custom layout at one point or another. An example of a CMOS inverter can be used to illustrate this procedure. (See Figs. A.11 and A.12.)

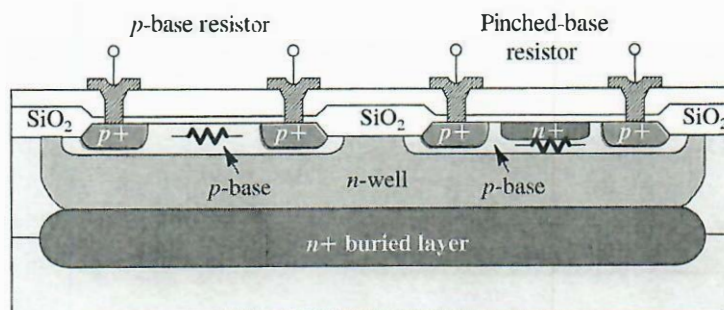


Fig. A.10 *p*-base and pinched-base resistors.

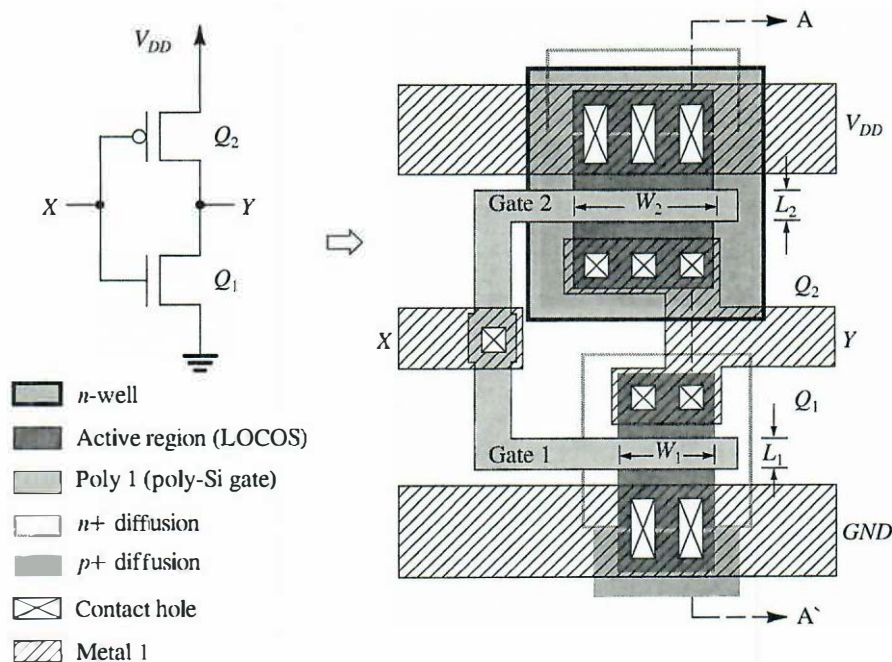


Fig. A.11 A CMOS inverter schematic and its layout.

The circuit must first be “flattened” and redrawn to eliminate any interconnection cross-overs, similar to the requirement of a printed circuit board layout. Each process is made up of a specific set of masking layers. In this case, seven layers are used. Each layer is usually assigned a unique color and fill pattern for ease of identification on a computer screen or on a printed color plot. The layout begins with the placement of the transistors. For illustration purposes, the p - and n -MOSFETs are placed in similar arrangement as the schematic. In practice, the designer is free to choose the most area-efficient layout. The MOSFETs are defined by the active areas overlapped by the “poly 1” layer. The MOS channel length and width are defined by the width of the “poly 1” strip and that of the active region, re-

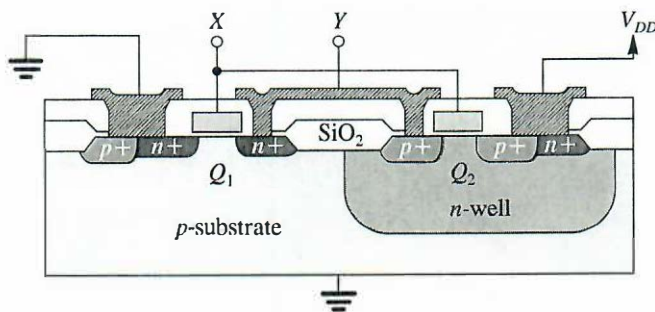


Fig. A.12 Cross section along the plane AA' of a CMOS inverter.

spectively. The p -MOSFET is enclosed in an n -well. For more complex circuits, multiple n -wells can be used for different groups of p -MOSFETs. The n -MOSFET is enclosed by the $n+$ diffusion mask to form the source and drain, whereas the p -MOSFET is enclosed by the $p+$ diffusion mask. Contact holes are placed in regions where connection to the metal layer is required. Finally, the "metal 1" layer completes the interconnections.

The corresponding cross-sectional diagram of the CMOS inverter along the AA' plane is as shown in Fig. A.12. The poly-Si gates for both transistors are connected to form the input terminal, X . The drains of both transistors are tied together via "metal 1" to form the output terminal, Y . The sources of the n - and p -MOSFET are connected to GND and V_{DD} , respectively. Note that butting contacts consist of side-by-side $n+/p+$ diffusions and are used to tie the body potential of the n - and p -MOSFETs to the appropriate voltage levels.

Once the layout is completed, the circuit must be verified using CAD tools such as a circuit extractor, design rule checker (DRC), and circuit simulator. Once these verifications are satisfied, the design can be "taped out" to a mask-making facility. A pattern generator (PG) machine can then draw the geometries on a glass or quartz photoplate using electronically driven shutters. Layers are drawn one by one on different photoplates. After these plates are developed, clear and dark patterns resembling the geometries on the layout will result. A set of the photoplates for the CMOS inverter example is shown in Fig. A.13. Depending on whether the drawn geometries are meant to be opened as windows or kept as patterns, the plates can be **clear** or **dark** fields. Note that each of these layers must be processed in sequence. They must be aligned within very fine tolerances to form the transistors and interconnections. Naturally, the greater the number of layers, the more difficult it is to maintain the alignment. This also requires better photolithography equipment and possibly results in lower yield. Hence, each additional mask will be reflected in an increase in the final cost of the IC chip.

SUMMARY

This appendix presented an overview of the various aspects of VLSI fabrication procedures. This includes component characteristics, process flows, and layouts. This is

by no means a complete account of state-of-the-art VLSI technologies. Interested readers should consult reference books on this subject for more detailed discussions.

BIBLIOGRAPHY

R. S. Muller and T. I. Kamins, *Device Electronics for Integrated Circuits*, 2nd ed., New York, John Wiley & Sons, 1986.

W. R. Runyan and K. E. Bean, *Semiconductor Integrated Circuit Processing Technology*, New York, Addison Wesley, 1990.

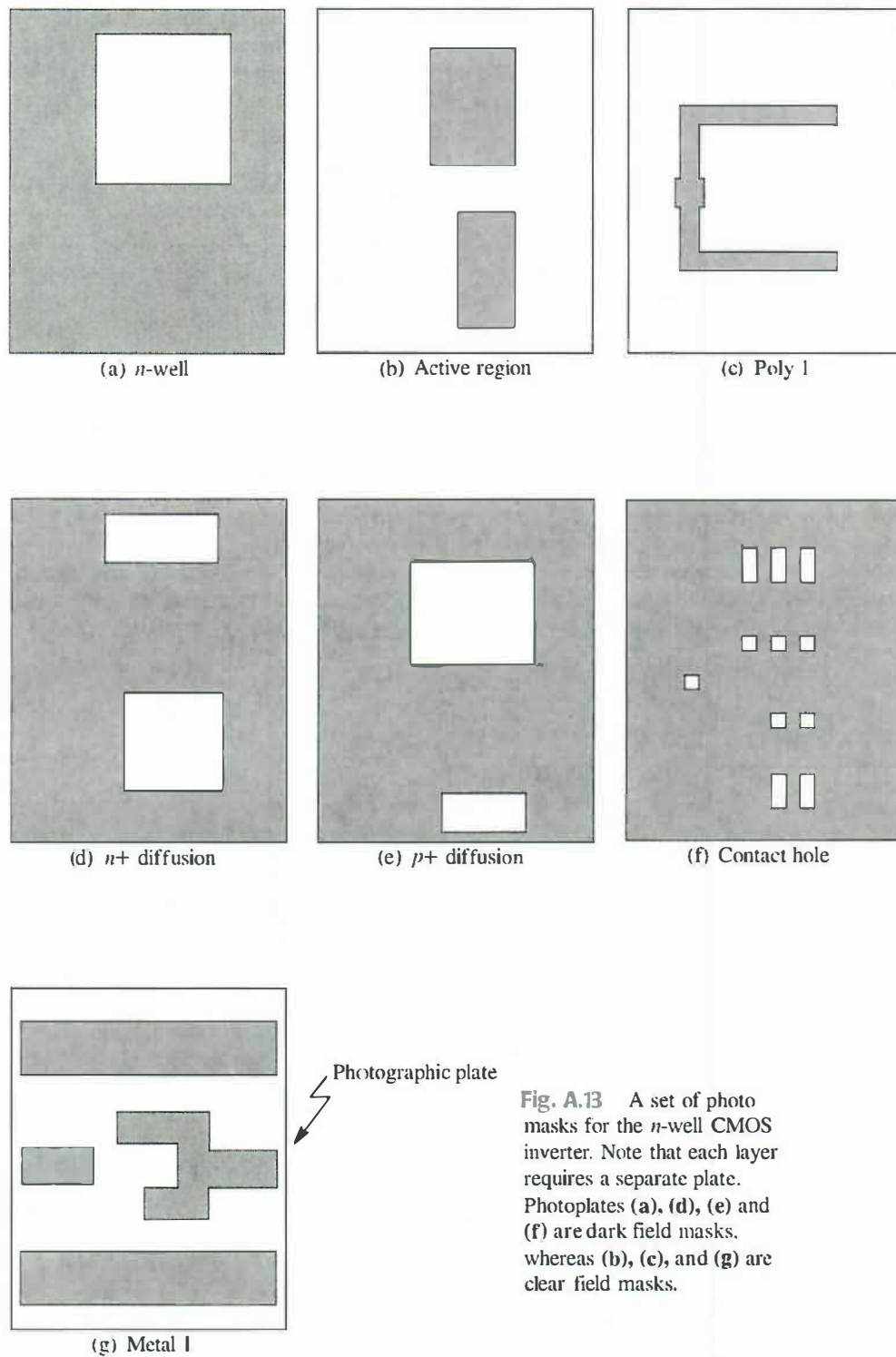


Fig. A.13 A set of photo masks for the n -well CMOS inverter. Note that each layer requires a separate plate. Photoplates (a), (d), (e) and (f) are dark field masks, whereas (b), (c), and (g) are clear field masks.